의미 유사도를 활용한 Distant Supervision 기반의 트리플 생성 성능 향상

윤희근○, 최수정, 박성배, 박세영
경북대학교 컴퓨터학부
{hkyoon,sjchoi,sbpark,sypark}@sejong.knu.ac.kr

Improving The Performance of Triple Generation Based on Distant Supervision By Using Semantic Similarity

Hee-Geun Yoon○, Su Jeong Choi, Seong-Bae Park, Se-Young Park
School of Computer Science and Engineering, Kyungpook National University

요 약

본 논문에서는 한국어 트리플 생성 시스템의 정확도를 향상시키기 위한 distant supervision 기반의 신뢰도 측정 방법을 제안한다. 기존의 많은 패턴 기반의 트리플 생성 시스템에는 distant supervision의 기본 가정으로 인해 다수의 오류 패턴이 발생할 여지가 크다. 기존의 연구에서는 오류 패턴을 제거하기 위하여 발생 빈도, 공기 횟수 등의 통계에 기반하여 간접적으로 신뢰도를 측정하였다. 본 논문에서는 한국어 패턴과 영어 프로퍼티 사이의 의미 유사도를 측정함으로써 통계에 기반한 방법보다 더 정확한 신뢰도 측정 방법을 제안한다. 비지도 학습 방법인 워드임베딩을 활용하여 어휘의 의미를 학습하고, 이들 사이의 유사도를 측정한다. 한국어 패턴과 영어 프로퍼티의 어휘 불일치 문제를 해결하기 위하여 정준상관분석을 활용하였다. 실험 결과에 따르면 본 논문에서 제안한 패턴 신뢰도 측정 방법은 통계 기반의 방법에 비해 정확률이 9%나 더 높은 트리플 집합을 생성함을 보여주어, 의미 유사도를 반영한 신뢰도 측정이 기존의 통계 기반 신뢰도 측정보다 고품질 트리플 생성에 더 적합함을 확인하였다.

주제어: Word Embedding, Distant Supervision, Semantic Similarity. Canonical Correlation Analysis

1. 서론

오늘날의 인터넷은 전 세계의 수많은 사용자가 작성한 대량의 정보가 구축되어 있으며, 날마다 새로운 정보가 추가, 갱신되고 있다. 이런 정보들은 다양한 웹 서비스를 통해 사용자에게 제공된다. 현재 인터넷의 대부분을 차지하고 있는 정보들은 비구조 자연어로 표현되어 있다. 자연어로 표현된 문서들은 일반 사용자들에게 매우 일반적으로 보편적인 표현법이다. 하지만 이런 비구조 표현은 컴퓨터가 활용하기에는 부적합하다.

최근에 비구조 표현으로 구축된 정보를 구조화하여 컴퓨터가 계산 가능한 형태로 구축하여 제공하는 서비스가 늘어나고 있다. Wikidata, DBpedia, YAGO 등 다양한 지식베이스가 구조화된 데이터를 제공하고 있다. 하지만 구조화된 데이터를 생성하는 일은 매우 큰 비용을 요구하는 일이기 때문에 인터넷에 있는 방대한 비구조 자료에 비하면 그 양이 매우 적다. 특히 이런 구조화된 데이터는 대부분 영어권을 기준으로 구축되고 있어, 구조화된 한국어 데이터는 거의 존재하지 않는 것이 현실이다.

구조화된 데이터 생성에 드는 비용 문제를 해소하기 위하여 자동으로 구조화된 데이터를 생성하는 지도, 반지도 학습 방식의 다양한 연구가 이루어지고 있다. 일반적인 자연어 문장으로부터 <주어, 프로퍼티, 목적어> 트리플 형태의 데이터를 추출하고 이를 이용하여 기존 지식베이스를 확장한다. 지도 학습 방법에 기반한 연구들에서는 한 문장에서 함께 나타난 객체들 쌍의 관계(프로퍼티)를 판별하는 문제로 정의하여 접근하였다. 자질 기반[1], 구조 표현[2] 등 다양한 모델들이 제안되었으며, 매우 높은 정확도를 보여주었다. 하지만 지도 학습 기반의 모델에서는 정답이 부착된 학습데이터를 요구하는데, 학습 데이터 구축에 매우 큰 비용이 요구되기 때문에 다양한 도메인에 적용하기에는 한계가 있었다. 이를 해결하기 위하여 반지도 방법들이 제안되었으며, 특히 최근에 distant supervision 기반의 모델들이 많이 제안되었다 [3,4,5].

Distant supervision은 반지도 학습 방법으로, 다음과 같은 가정 하에서 모델을 학습한다. 이 가정은 특정 트리플의 주어와 목적어 객체를 포함하고 있는 문장은 주어진 트리플 프로퍼티의 의미를 표현하고 있다는 것이다. 이로 인해 모델을 학습하고자 하는 프로퍼티의 일부 시드 트리플과 코퍼스만 있으면 별도의 학습 데이터 구축이 없어도 학습이 가능하다. 기존의 많은 연구에서는 이렇게 추출된 문장에 포함된 어휘, 품사 태그 등을 이용하여 정의한 패턴 형태를 활용하였다 [3,4].

Distant supervision은 적용의 편리성으로 인해 다양한 연구에서 사용되었지만 한 가지 단점이 존재한다. 트리플의 두 객체를 포함하고 있는 문장이 트리플 프로퍼티의 의미를 표현하고 있을 것이라는 가정이 항상 옳지는 않다는 것이다. 예를 들어, 트리플 <*글로리아 스튜어트, birthPlace, 캘리포니아 주*>이 주어지면 distant supervision 가정에서는 다음의 문장에서 패턴을 추출한

다.

"글로리아 스튜어트는 폐암으로 투병하던 중 2010년 9월 26일 캘리포니아 주의 자택에서 사망했다."

하지만 이 문장은 글로리아 스튜어트가 캘리포니아 주에서 사망했음을 의미하고 있는 문장으로, 주어진 트리플의 프로퍼티인 *birthPlace*와는 전혀 다른 의미를 표현하고 있다. 그렇기 때문에 이 문장으로부터 생성된 패턴은 *birthPlace* 관계를 가지는 새로운 트리플 생성에 사용하기에는 부적합하다. 패턴은 대상 프로퍼티의 의미를 잘 나타낼 수 있어야하기 때문에, 위의 예와 같이 대상 프로퍼티의 의미를 나타내지 못하는 오류 패턴은 제거되어야 한다. 오류 패턴을 제거하기 위하여 기존의 연구에서는 빈도 또는 프로퍼티와 공기하는 통계 정보를 활용한 신뢰도 측정 방식을 사용하였다[3,4,5]. 하지만 통계 정보는 대상 프로퍼티와 패턴 사이의 관계성을 간접적으로 추론하는 것일 뿐, 신뢰도 측정에 고려되어야 할 패턴과 프로퍼티의 의미적인 관계성을 측정하지는 못한다. 그렇기 때문에 비록 이 방법이 대량의 데이터에서 유의미한 결과를 제공하더라도 한계가 있다. 또한 데이터가 충분하지 않을 경우에는 신뢰도를 전혀 측정할 수 없다.

본 논문에서는 distant supervision 기반의 한국어 트리플 생성 시스템의 성능 향상 방법을 제안한다. Distant supervision 기반에서 성능 향상을 위한 핵심은 시드로 주어진 트리플의 의미를 포함하지 않은 문장에서 생성된 패턴을 잘 제거하는 것이다. 일반적으로 문장으로부터 생성된 패턴의 경우, 자연어 어휘로 구성되어 있고 트리플의 프로퍼티 또한 레이블(Label) 또는 식별자(Identifier) 등의 속성을 통해서 사람이 이해 가능한 어휘들로 그 의미를 표현하고 있다. 본 논문에서는 이들을 활용하여 패턴과 대상 프로퍼티와의 의미 유사도를 신뢰도 측정에 사용함으로써 기존의 빈도 기반의 신뢰도 측정보다 더 정확한 신뢰도를 측정한다. 의미 유사도 측정을 위해서 최근 다양한 연구에서 우수한 성능을 보여주는 비지도 학습 방법인 워드임베딩 방법을 채택함으로써 큰 비용을 들이지 않고도 높은 정확도의 의미 유사도를 측정한다.

제안한 시스템이 적용할 트리플은 패턴과 프로퍼티의 언어가 다르다. 패턴은 한국어 문장으로부터 생성되기 때문에 한국어로 구성되어 있는 반면 대부분의 프로퍼티들은 영어로 표현되어 있기 때문에, 패턴과 프로퍼티의 워드임베딩 공간은 독립적으로 학습되며 이로 인해 이들의 의미적인 유사도를 직접적으로 측정하기 힘들다. 이를 해결하기 위하여 독립적으로 학습된 워드임베딩 공간들을 정준상관분석(Canonical Correlation Analysis)을 이용하여 동일한 저차원으로 투영함으로써 서로 다른 언어로 구성된 패턴과 프로퍼티의 유사도를 계산할 수 있도록 한다.

디비피디아와 위키피디아를 사용하여 수행한 실험 결과에 따르면 기존에 많이 사용되는 통계 기반 신뢰도 측정 방법보다 본 논문에서 제안하는 의미 유사도 기반의 신뢰도 측정법이 더 정확한 트리플 생성에 기여함을 확인하였다. 각 신뢰도 값에 따라 상위 2,000개의 트리플을 평가한 결과, 통계 기반 방법보다 제안한 방법이 더 우수한 성능을 보여주었다.

2. 관련 연구

최근 많은 연구를 통해 distant supervision에 기반한 구조화된 데이터 생성이 제안되었다. Gerber et al.[3]은 영어 문서에서 distant supervision에 기반하여 어휘 패턴을 생성하고, 이를 통해 새로운 트리플을 생성하는 BOA를 제안하였다. BOA는 주어 객체와 목적어 객체가 포함된 문장에서 두 객체 사이에 존재하는 어휘들을 추출하여 패턴으로 사용하였다. 신뢰도 추출을 위하여 support, typicity 그리고 specicity라는 3가지 통계 값에 기반한 신뢰도 측정 함수를 사용하였다. Wu et al.[4] 역시 distant supervision 기반의 관계 추출 시스템 WOE를 제안하였다. WOE는 BOA와 다르게 문장의 의존 관계 트리로부터 패턴을 생성하여 먼 거리 의존 관계(Long dependency problem) 문제를 해결하였다. 하지만 WOE 역시 패턴의 신뢰도 측정을 위하여 단순 통계 값을 활용하였다. 이현구 외[5]는 suffix tree와 distant supervision에 기반한 관계 추출 방법을 제안하였다. WOE와 유사하게 의존 관계 트리를 활용하여 관계 추출 규칙을 자동으로 생성하였다. 이 논문에서도 문장을 통해 추출된 규칙의 점수를 빈도에 기반한 통계 값으로 측정하였다.

기존의 많은 연구에서 distant supervision에 기반하여 다양한 도메인에서 우수한 실험 결과를 보여주었다. 하지만 이런 연구들은 오류 패턴을 제거하기 위하여 통계에 기반한 신뢰도 함수를 사용하였으며, 이 함수들은 패턴의 적합도를 판단하는데 있어 기준이 되는 패턴과 프로퍼티의 의미 관계를 직접적으로 측정하지 못한다는 한계가 있다. 그렇기 때문에 이 신뢰도 측정 방법을 개선한다면 성능을 더 향상 시킬 여지가 존재한다. 그렇기 때문에 패턴의 신뢰도를 정확하게 측정하기 위해서는 통계 정보와 같은 간접적인 방법이 아니라 패턴과 프로퍼티의 의미적인 관계성을 직접적으로 측정할 수 있는 방법이 필요하다.

3. 의미 유사도를 이용한 트리플 생성 정확도 향상

3.1. 시스템 구조

본 논문에서는 그림 1과 같은 과정을 거쳐 패턴 생성, 패턴 필터링, 새로운 트리플 생성 과정을 수행한다. 주어진 지식베이스의 트리플과 코퍼스를 이용하여 새로운 트리플 추출에 사용하기 위한 각 프로퍼티 별 패턴 후보를 학습한다. 이 단계에서는 특정 트리플의 주어와 목적어 개체를 포함하고 있는 문장은 주어진 트리플 프로퍼티의 의미를 표현할 것이라는 distant supervision 가정을 이용하여, 주어진 트리플의 주어와 목적어 객체에 해당하는 문장들을 추출하고, 이 문장들을 이용하여 패턴 후보를 추출한다.

본 시스템에서 학습된 패턴들은 distant supervision 가정의 한계로 인해 다수의 오류가 포함되어 있을 수 있으며 이로 인해 최종적인 트리플 정확도의 저하를 유발

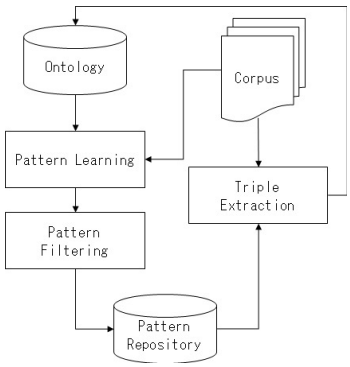

그림 1. 지식 베이스 확장을 위한 시스템 구성

한다. 이를 해결하기 위하여 패턴 필터링 과정을 거쳐 오류 패턴을 제거한다. 최종적으로 필터링된 패턴과 코퍼스를 이용하여 새로운 트리플을 생성하고 이를 통해 지식베이스를 확장한다.

3.2. 새로운 트리플 추출을 위한 패턴 학습

기존 연구는 문장에서 새로운 트리플 추출을 위한 패턴 생성 방법을 제안하였다. 일반적으로 트리플의 프로퍼티에 해당하는 부분은 문장에서 술어로 표현되기 때문에 기존의 영어권 연구에서는 주어진 트리플의 주어, 목적어 객체 사이에 포함되어 있는 어휘들을 이용하여 패턴을 생성하였다 [3]. 하지만 한국어는 영어와 달리 술어가 주어, 목적어 사이에 위치하지 않기 때문에 주어, 목적어 사이의 어휘를 사용하는 것은 부적절하다. 또한 한국어는 영어와 달리 어순이 자유롭기 때문에 두 객체 사이의 관계를 표현하는 술어는 찾기에 어려움이 있다.

제안한 방법에서는 한국어 문장에서 패턴을 추출하기 위하여 윤희근 외[6]가 제안한 의존 관계 트리에 기반한 패턴 생성 방식과 유사한 방식을 사용한다. 이 연구에서는 주어와 목적어 어휘 노드들 사이의 패스에 존재하는 술어만을 패턴으로 사용하였으나, 본 논문에서는 두 노드 사이에 존재하는 모든 어휘를 패턴으로 사용한다. 이를 통해 두 객체 사이의 관계를 표현하는 어휘들만을 추출하고, 그 이외의 어휘들은 쉽게 제거할 수 있다. 그림 2는 예제 문장 **"글로리아 스튜어트는 폐암으로 투병하던 중 2010년 9월 26일 캘리포니아 주의 자택에서 사망했다."** 로부터 생성된 패턴의 예를 보여준다.

3.3. 의미 유사도 측정을 위한 워드임베딩 공간

Distant supervision 가정 기반의 자가 지식 학습 과정에서 가장 중요한 것은 주어진 트리플의 의미를 포함하지 않은 문장으로부터 생성된 패턴을 효과적으로 필터링하는 것이다. 기존 연구들은 오류 패턴을 제거하기 위하여 패턴의 발생 빈도, 프로퍼티와의 공기 횟수 등 통계 정보를 활용하여 신뢰도를 측정하였다. 하지만 통계 정보는 신뢰도 측정에 고려되어야 할 패턴과 프로퍼티의 의미적인 관계성을 직접적으로 측정할 수 없다는 한계가 있다.

본 논문에서는 통계 기반 신뢰도 함수의 한계를 극복하기 위하여 패턴과 대상 프로퍼티의 의미 유사도를 직

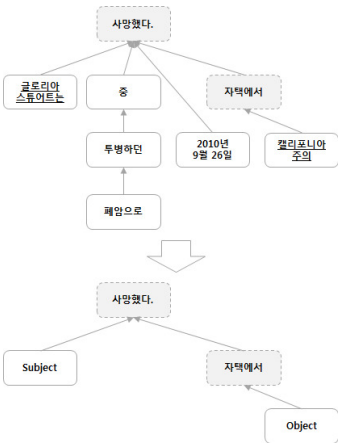

그림 2. 의존 관계 트리 기반 패턴 생성 예제

접적으로 측정하여 신뢰도로 사용하는 방법을 제안한다. 앞서 언급한 것처럼 패턴은 문장에서 추출된 어휘의 집합으로 의미를 표현하고 프로퍼티 또한 레이블, 식별자 등의 속성을 통해 의미를 나타낸다. 그러므로 이 어휘들 사이의 의미 유사도를 측정하면 각 패턴이 대상 프로퍼티에 대해서 얼마나 적합한지를 측정할 수 있다. 패턴 p와 프로퍼티 r의 의미 유사도에 기반한 신뢰도는 아래와 같이 정의된다.

$$confidence(p,r) = sim_{semantic}(p,r)$$

어휘의 의미 유사도 측정을 위해서는 많은 방법이 제안되었으나, 최근에는 대규모 데이터를 활용한 비지도 학습 방법인 워드임베딩 방식이 우수한 성능을 보여주고 있다. 워드임베딩은 어휘를 N차원의 벡터공간에 맵핑하여 분산 표상(Distributed Representation)으로 표현하는 방법이며, 워드임베딩 공간이란 어휘들이 맵핑되어 있는 공간을 의미한다. Mikolov et al.[7]은 대량의 코퍼스를 이용하여 단어들 사이의 공기 정보를 활용하면 학습된 워드임베딩 공간에서 의미적으로 유사한 단어가 비슷한 형태의 벡터로 학습되는 것을 실험적으로 보였다. 이를 통해 어휘들 사이의 의미 유사도를 워드임베딩 공간에 맵핑된 각 어휘들의 벡터 유사도를 구하는 것으로 측정할 수 있다. 워드임베딩을 통한 패턴과 프로퍼티의 의미 유사도는 각각의 워드임베딩 공간에서의 벡터 표현 $\vec{p}$, $\vec{r}$의 유사도로 아래와 같이 구해진다.

$$sim_{semantic}(p,r) = sim_{we}(\vec{p},\vec{r})$$

벡터 유사도는 아래와 같이 코사인 유사도로 계산할 수 있다.

$$sim_{we}(\vec{p},\vec{r}) = cos(\vec{p},\vec{r}) \quad (1)$$

패턴과 프로퍼티가 각각 하나의 어휘로 구성되어 있을 경우에는 위와 같이 두 어휘 벡터의 코사인 유사도를 계산한다. 하지만 패턴과 프로퍼티는 둘 이상의 어휘로 구성될 수도 있다. 예를 들어 패턴은 여러 어절로 구성될

수 있으며, 프로퍼티의 경우 한 어절로 표현되지만 *hasChild* 와 같이 두 어휘 이상이 결합된 경우도 있다. 워드임베딩은 어휘별로 벡터를 학습하기 때문에 두 단어 이상이 결합된 어휘의 경우 바로 벡터로 표현할 수 없다. 이에 본 논문은 두 개 이상의 어휘로 구성된 패턴과 프로퍼티의 벡터를 구성 요소들의 평균 벡터로 정의하였다. 예를 들어 패턴 p가 n개의 어휘들로 구성되어 $p=\{pw_1,...,pw_n\}$이고 프로퍼티 r이 $r=\{rw_1,...,rw_m\}$와 같이 m개의 어휘로 구성될 때, 패턴 및 프로퍼티의 벡터는 다음과 같이 정의된다.

$$\mathbf{pw}=\frac{1}{n}\sum_{pw\in p}\overrightarrow{pw} \quad (2)$$

$$\mathbf{rw}=\frac{1}{m}\sum_{rw\in r}\overrightarrow{rw} \quad (3)$$

수식 2과 3을 통해 패턴과 프로퍼티의 벡터를 구할 수 있고, 수식 1의 패턴과 프로퍼티 코사인 유사도는 아래와 같이 계산한다.

$$sim_{we}(p,r)=\frac{\mathbf{pw}\cdot\mathbf{rw}}{\|\mathbf{pw}\|\times\|\mathbf{rw}\|}$$

3.4. 이종 언어 워드임베딩 벡터의 유사도 계산

앞서 보았듯이 한국어 문장에서 생성되는 패턴은 한국어 어휘들로 구성되어 있는 반면 디비피디아나 위키데이터와 같이 널리 활용되고 있는 지식베이스의 프로퍼티는 대부분 영어로 표현되어 있다. 이로 인해 패턴의 경우는 한국어 코퍼스로부터 학습된 워드임베딩 공간이 필요하고, 프로퍼티는 영어 코퍼스로부터 학습된 영어 워드임베딩 공간을 필요로 한다. 이 두 공간은 서로 독립적인 코퍼스로부터 학습되기 때문에 최종적으로 학습된 워드임베딩 공간 또한 서로 독립적이다. 또한 이 두 공간은 차원 수가 다르게 학습될 수도 있고, 차원 수가 동일하다 하더라도 같은 의미를 가진 한국어, 영어 어휘가 완전 다른 벡터로 학습될 수도 있다. 그림 3은 독립적으로 학습된 한국어와 영어 워드임베딩 공간에서 동일한 의미를 가지는 어휘 3쌍의 벡터를 보여준다. 그림 3에 의하면 서로 동일한 의미를 가지는 어휘들이 완전 다른 형태의 벡터로 학습된 모습을 확인 할 수 있다.

위와 같이 독립적으로 학습된 워드임베딩 공간에서 각각 추출되는 패턴과 프로퍼티 벡터로는 둘 사이의 의미 유사도를 제대로 측정할 수 없다. 본 논문에서는 이 문제를 해결하기 위해 Faruqui et al.[8]이 제안한 정준상관분석에 기반한 이종 언어 워드임베딩 공간 투영 방식을 사용한다. 미리 주어진 동일한 의미를 가진 어휘 쌍들을 이용하여 이들의 상관계수가 높아지게 하는 투영 행렬을 학습함으로써 이종 공간의 벡터들을 동일한 공간으로 투영시킨다. 이렇게 구해진 투영 공간에서는 동일한 의미를 가진 이종 언어 벡터들이 비슷한 형태로 나타나기 때문에 다른 언어의 어휘 유사도를 측정할 수 있다.

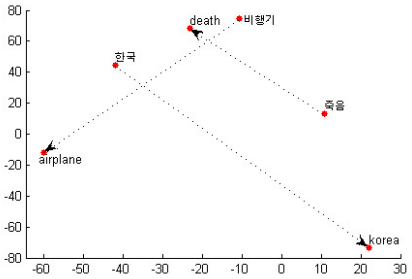

그림 3. 독립적으로 학습된 워드임베딩 공간 결합 예

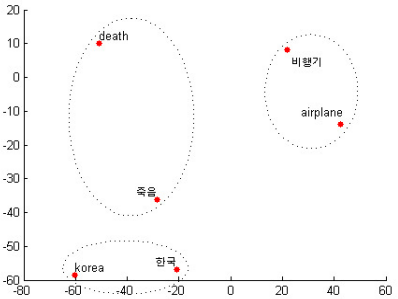

그림 4. CCA를 통해 투영된 워드임베딩 공간

$\mathbb{K}\in\mathbb{R}^{n_1\times d_1}$, $\mathbb{E}\in\mathbb{R}^{n_2\times d_2}$ 는 각각 한국어 및 영어 코퍼스로부터 학습된 워드임베딩 공간의 벡터이고 n_1, n_2는 각각 한국어, 영어 어휘 수이며 d_1, d_2는 각 워드임베딩 공간의 차원 수이다. 여기서 미리 주어지는 동일한 의미를 가지는 이종 언어의 어휘 쌍만을 포함하는 $\mathbb{K}'\in\mathbb{R}^{n\times d_1}$, $\mathbb{E}'\in\mathbb{R}^{n\times d_2}$가 있을 때, 정준상관분석을 통해서 아래와 같이 투영 행렬 U, V를 구할 수 있다.

$$U,V=\mathrm{CCA}(\mathbb{K}',\mathbb{E}')$$

정준상관분석을 통해 구해지는 행렬 $U\in\mathbb{R}^{d_1\times d}$, $V\in\mathbb{R}^{d_2\times d}$는 한국어, 영어 워드임베딩 공간의 벡터들을 동일한 d차원 공간으로 투영시켜주는 역할을 한다. 이 행렬들을 통해 독립적으로 학습된 전체 한국어 워드임베딩 $\mathbb{K}$와 영어 워드임베딩 $\mathbb{E}$를 다음과 같이 공통된 저차원으로 투영할 수 있다.

$$\mathbb{K}^*=\mathbb{K}V$$

$$\mathbb{E}^*=\mathbb{E}W$$

그림 4는 이 과정을 통해 투영된 이종 언어 워드임베딩 공간을 보여준다. 그림 4의 워드임베딩 공간은 그림 3의 독립된 워드임베딩 공간과 달리 동일한 의미를 가진 어휘들이 가깝게 배치되어 있음을 볼 수 있다. 저차원으로 투영된 패턴과 프로퍼티 벡터들의 유사도 측정을 위하여 수식 2, 3은 아래와 같이 변경한다.

$$\mathbf{pw}=\frac{1}{n}\sum_{pw\in p}\overrightarrow{pw}\,V$$

$$\mathbf{rw}=\frac{1}{m}\sum_{rw\in r}\overrightarrow{rw}\,W$$

4. 실험

4.1. 데이터 셋

본 논문에서는 제안한 방법의 성능을 보이기 위하여 3.1의 시스템을 구현하여 패턴 학습 및 트리플 생성을 수행하였다. 시드 지식베이스로 한국어 디비피디아를 이용하였고, 패턴 생성 및 새로운 트리플 생성에 사용할 코퍼스로 한국어 위키피디아를 코퍼스로 사용하였다. 위키피디아의 자연어처리를 위하여 ETRI 언어 분석기를 사용하였다. 표 1은 패턴 및 트리플 생성 실험에 사용된 지식베이스와 코퍼스의 간략한 통계 자료이다.

표 1. 트리플 생성을 위한 데이터 통계

항 목	값
디비피디아 트리플	513,170
유일한 프로퍼티	449
한국어 위키피디아 문장수	2,862,172

제안한 의미 유사도를 측정하기 위하여 한국어 및 영어 워드임베딩 공간의 학습이 필요하다. 이를 위하여 한국어와 영어 위키피디아를 이용하여 각 언어별 워드임베딩 공간을 학습하였다. 워드임베딩 공간 학습은 오픈 소스인 word2vec[9]를 이용하여 수행하였다. 두 언어 워드임베딩의 정렬을 위하여 초기에 주어질 동일 의미 어휘쌍은 Faruqui et al.[8]에서 사용한 방식과 동일하게 수행하였다. 이는 WIT^3[10]의 한국어-영어 병렬 코퍼스를 사용하여 자동으로 추출하였다. 표 2는 워드임베딩 학습에 사용된 코퍼스 및 학습된 워드임베딩 공간의 간략한 통계를 보여준다.

표 2. 워드임베딩 공간 학습을 위한 데이터 통계

항목	값
영문 위키피디아 어휘 수	3,000,000,000 이상
한국어 위키피디아 어휘 수	79,408,956
유일한 한국어 어휘 수	190,367
유일한 영어 어휘 수	489,424
동일 의미 어휘 쌍	10,048
워드임베딩 학습 차원	200

패턴과 프로퍼티가 두 단어 이상일 경우 이를 분리하는 작업이 필요하다. 이를 위하여 한국어의 경우 띄어쓰기 단위로 구분하였으며, 프로퍼티의 경우는 미리 정의한 단순한 룰에 기반하여 어휘들을 분리하였다.

4.2. 비교 모델 및 성능 측정 방법

본 논문에서 제안한 의미 유사도를 활용한 신뢰도 측정 방법의 우수성을 보이기 위하여, 기존 연구에서 사용된 통계 정보 기반의 신뢰도 측정 방법과 비교하였다. 통계 정보 기반의 신뢰도는 프로퍼티별로 각 패턴의 발생 빈도를 사용하였다. 신뢰도 측정 이외에 패턴 생성 및 트리플 생성 과정은 모두 동일하게 수행하였다.

디비피디아와 위키피디아 전체 데이터를 이용하여 트리플을 생성할 경우, 그 양이 너무 많아 모든 트리플을 평가하기가 어렵다. 본 논문에서는 두 신뢰도 함수의 성능 비교를 위해서 Top-K 정확도를 측정하였다. 이를 위하여, 우선 패턴 필터링을 수행하지 않은 상태에서 모든 패턴으로 트리플을 생성한다. 그리고 각각의 신뢰도 측정 함수를 이용하여 모든 패턴의 신뢰도를 측정한 후, 패턴을 신뢰도 값을 기준으로 정렬한다. 이렇게 각각의 신뢰도로 정렬된 패턴들에 대해서, 각 패턴들이 생성한 트리플들을 상위 K 개를 추출한다. 이렇게 추출된 트리플들에 대해서 사실 여부를 체크하여 수작업으로 정확도를 측정하였다.

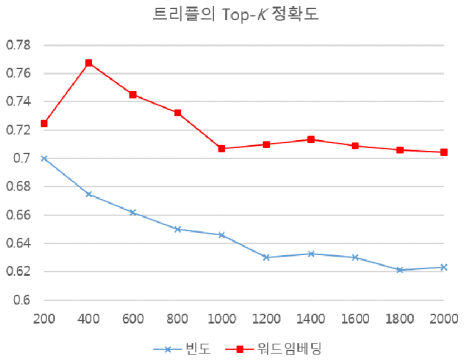

그림 5. 생성된 트리플의 Top-K 정확률

4.3. 실험 결과

디비피디아와 한국어 위키피디아를 대상으로 총 25,784개의 패턴과 422,733개의 트리플을 생성하였다. 두 방법의 성능 비교를 각 신뢰도 방법으로 측정된 데이터에서 상위 2,000개씩의 트리플을 추출하여, 이들을 대상으로 정확도를 수작업 평가하였다. 그림 5는 Top-K 트리플의 정확도를 보여준다. 그래프에 따르면 제안한 방법이 빈도 기반의 신뢰도 측정 방법에 비하여, Top-200을 제외하고 모든 구간에서 약 10% 정도의 성능 차이를 보여주었다. 최종적으로 상위 2,000개의 트리플을 추출하였을 때, 본 논문에서 제안한 방법은 약 71%의 정확도를 보여주는데 반해, 빈도에 기반한 신뢰도 측정 방법은 약 62%의 성능을 보여주어 9%의 성능 차를 보여주었다.

Top-200의 성능 저하 원인은 워드임베딩 공간에서 학습된 의미 관계가 항상 유의어가 아니기 때문인 것으로 판단된다. 적합한 패턴이란 프로퍼티의 의미를 잘 나타내는 패턴이므로 유의어의 관계를 가지는 어휘들로 표현되어야 한다. 일반적으로 코퍼스 기반의 방법들은 유의어 관계를 찾는데 우수한 성능을 보여준다 [11]. 하지만 코퍼스에 기반한 워드임베딩 학습 방식은 문맥에 따라

전혀 다른 임베딩 공간이 학습될 수 있으며, 이로 인해 어휘들 사이의 유사도가 전혀 다르게 유도될 수도 있다 [12]. 실제로 학습된 영어 워드임베딩을 살펴보면 *spouse*와 *grandparent*가 매우 유사도 값을 갖도록 학습된 것을 확인할 수 있었다. 이 두 어휘는 가족이라는 어휘의 유사 어휘 집합으로 해석되어 높은 유사도를 가질 수도 있지만, 사실 전혀 다른 의미를 가진 어휘이다. 워드임베딩의 이런 특징들로 인해 제안한 방법의 성능이 저하된 것으로 판단된다.

실험 결과에 따르면 워드임베딩에 기반한 유사도 측정에 한계가 있음에도 불구하고 더 정확한 트리플을 생성하는데 큰 기여를 함을 볼 수 있었다. 이를 통해 패턴의 신뢰도 측정에 패턴과 프로퍼티의 직접적인 의미 유사도를 반영하는 것이 통계 기반의 간적접인 측정 방법에 비해 더 적합함을 알 수 있다.

5. 결론

본 논문에서는 distant supervision 가정에 기반한 트리플 생성 시스템의 성능을 향상 시킬 수 있는 방법을 제안하였다. Distant supervision은 반지도 학습 방식으로 학습 데이터 구축에 드는 비용이 적기 때문에 다양한 도메인에 쉽게 적용할 수 있다. 하지만 이 방법의 가정의 한계로 인하여 잘못된 패턴이 생성될 수 있으며, 이는 최종적으로 생성하는 트리플의 품질을 저하시킬 수 있다. 본 논문에서는 오류 패턴으로 인한 성능 저하를 해소하기 위하여 의미 유사도를 활용한 패턴 신뢰도 측정 방법을 제안하였다. 워드임베딩에 기반한 패턴과 프로퍼티 사이의 의미 유사도를 측정함으로써 패턴의 신뢰도를 정확하게 측정할 수 있다. 하지만 본 논문에서는 이종의 언어로 구성된 패턴과 프로퍼티를 고려하기 때문에 이들 사이의 유사도를 직접적으로 측정할 수 없다. 이를 해결하기 위하여, 독립적으로 학습된 각 워드임베딩 공간을 CCA를 이용하여 공통된 새로운 워드임베딩 공간으로 투영시킴으로써 이 문제를 해결하였다.

실험 결과에 따르면, 제안한 의미 유사도 기반의 패턴 신뢰도 측정 방법은 기존에 많은 연구에서 사용된 통계 기반의 신뢰도 측정 방법에 비하여 트리플 생성 정확도 향상에 크게 도움이 됨을 보였다. 패턴의 신뢰도 기준으로 상위 2,000개의 트리플 정확도를 비교 하였을 때 본 논문에서 제안한 신뢰도 측정 방법이 통계 기반의 방법보다 9% 높은 정확도를 보였다.

향후 연구에서는 좀 더 정확한 의미 유사도 측정 방법이 필요할 것으로 생각된다. 본 논문에서 사용한 워드임베딩 방식은 비지도 학습 방식으로 손쉽게 적용할 수 있다는 장점이 있지만, 여전히 완벽한 의미 유사도를 측정하지는 못한다. 실제 실험에서도 이런 오류들이 성능 저하를 유발함을 확인하였다. 이를 해결하기 위하여 차후 연구로는 의미 유사도를 더 정확하게 측정할 수 있는 모델을 연구함으로써 패턴의 신뢰도를 더 정확하게 측정하고, 이를 통해 생성 트리플의 정확도를 더 향상 시킬 수 있는 방법을 연구할 예정이다.

감사의 글

이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원(No.R0101-15-0054, WiseKB : 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)과 2015년도 교육부 및 한국연구재단의 BK21 플러스 사업으로 지원을 받아 수행된 연구임 (No. 21A20131600005, 경북대학교 컴퓨터학부 Smart Life실현을 위한 SW인력양성사업단)

참고문헌

[1] G. Zhou, J. Su, J. Zhang, M. Zhang, "Exploring various knowledge in relation extraction", *Proceedings of the 43rd annual meeting on association for computational linguistics*, pp.427-434, 2005.

[2] A. Culotta, J. Sorensen, "Dependency Tree Kernels for Relation Extraction", *Proceedings of the 42nd annual meeting on association for computational linguistics*, 2004.

[3] D. Gerber, A.-C. Ngonga Ngomo, "Bootstrapping the linked data web", *Proceddings of the 1st Workshop on Web Scale Knowledge Extraction*, 2011.

[4] F. Wu, D. S. Weld, "Open information extraction using Wikipedia", *Proceddings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp.118-127, 2010.

[5] 이현구, 최맹식, 김학수, "Suffix Tree와 Distant Supervision을 이용한 관계 추출", 한글 및 한국어 정보처리 학술대회, pp.149-152, 2014.

[6] 윤희근, 박성배, "한국어 자가 지식 학습을 위한 패턴 및 인스턴스 생성", 한국지능시스템학회 논문지, 제25권, 제1호, pp.63-69, 2015.

[7] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality", *Proceedings of the Conference on Advances in Neural Information Processing Systems*, pp.3111-3119, 2013.s

[8] M. Faruqui, C. Dyer, "Improving Vector Space Word Representations Using Multilingual Correlation", *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp.462-471, 2014.

[9] https://code.google.com/p/word2vec/

[10] https://wit3.fbk.eu/

[11] Y. Chen, B. Perozzi, R. Al-Rfou, S. Skiena, "The expressive power of word embeddings", *Proceedings of the 30th International Conference on Machine Learning*, 2013.

[12] O. Levy, Y. Goldberg, "Dependency-based word embeddings", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp.302-308, 2014.