상위어 시퀀스의 클러스터링을 이용한 단어의 의미 애매성 해소

정창후*, 최윤수*, 최성필*, 윤화묵*
*한국과학기술정보연구원
e-mail:chjeong@kisti.re.kr

Word Sense Disambiguation using Hypernym Sequence Clustering

Chang-Hoo Jeong*, Yun-Soo Choi*, Sung-Pil Choi*, Hwa-Mook Yoon*
*Korea Institute of Science and Technology Information

요 약

본 논문에서는 과학기술문서에 존재하는 기술용어와 이들 간의 연관관계를 설명하는 디스크립터를 찾아서 [subject predicate object] 형태의 트리플을 생성하는 애플리케이션을 개발할 때 발생하는 단어 의미 애매성 해소 문제를 다룬다. 기술용어가 가지고 있는 연관관계를 결정하기 위해서 워드넷의 신셋 정보를 사용하는데 이 방법은 동사를 워드넷에 매핑할 때와 상위어 관계로 전이할 때 여러 개의 의미에 매핑되는 문제점이 발생한다. 이것을 해결하기 위해서 상위어 시퀀스 클러스터링을 이용한 단어의 의미 애매성 해결 방안을 제시한다. 이 방법을 사용함으로써 워드넷 매핑과 상위어 전이 시에 발생하는 다중 매핑 문제를 동시에 해결할 수 있다.

1. 서론

워드넷(WordNet)[1, 2]이 체계적으로 구축되면서 워드넷을 이용한 다양한 응용 시스템들이 개발되고 있다. 그러나 언어가 가지고 있는 특징으로 인해 워드넷을 각종 애플리케이션에 적용하기 위해서는 단어의 의미 애매성 문제를 해결해야 한다. 현재 많은 연구자들이 자신의 애플리케이션에서 발생하고 있는 단어의 의미 애매성 해소를 위해서 많은 노력을 기울이고 있다.

2. 관련연구

단어의 의미 애매성 해소 문제는 단순히 해당 단어만을 보고는 해결할 수 없기 때문에 여러 가지 정보를 활용해야 한다. 단어의 의미 애매성 해소를 위해서 통계적 수치를 이용하는 방법, 언어의 문법적 특성을 이용하는 방법, 외부의 지식 자원을 활용하는 방법 등의 다양한 방법론이 연구되어 왔다[3].

본 연구에서는 과학기술문서에 존재하는 기술용어와 이들 간의 연관관계를 설명하는 디스크립터를 찾아서 [subject predicate object] 형태의 트리플을 생성하는 애플리케이션[4]을 개발할 때 발생하는 단어 의미 애매성 해소 문제를 다룬다. 이때 predicate는 단순히 문장에 존재하는 디스크립터를 지칭하는 것이 아니라 해당 디스크립터로 표현되는 동사를 워드넷에 매핑한 후에 동사의 가장 상위 개념을 선택해서 연관관계로 할당하는 것이다. 따라서 해당 동사의 가장 추상화된 개념을 선택하는 문제라고 볼 수 있다. 이때 두 종류의 매핑 문제를 해결해야 하는데, 첫 번째는 디스크립터를 워드넷의 동사 신셋(synset)에 매핑할 때 처음에 어떤 신셋으로 매핑할 것인지를 결정해야 하는 문제이고, 두 번째는 선택된 신셋을 이용하여 가장 상위 개념의 신셋으로 개념 추상화를 수행할 때 어떤 상위어(hypernym)로 매핑할 것인지를 결정해야 하는 문제이다. 따라서 디스크립터의 추상화된 연관관계를 결정하기 위해서는 위의 두 가지 경우에 발생하는 매핑 문제를 모두 해결할 수 있어야 한다. 본 논문에서는 이러한 문제를 해결하기 위해서 상위어 시퀀스(hypernym sequence) 클러스터링을 이용한 단어의 의미 애매성 해결 방안을 제시한다.

3. 상위어 시퀀스 클러스터링을 이용한 디스크립터의 의미 결정

본 연구에서는 기술용어가 가지고 있는 연관관계를 결정하기 위해서 워드넷의 신셋 정보를 사용한다. 그러나 이전에 언급한 바와 같이, 이러한 방법은 동사를 워드넷에 매핑할 때와 상위어 관계로 전이할 때 여러 개의 의미(sense)에 매핑될 수 있기 때문에 의미 애매성 해결이 필요하다.

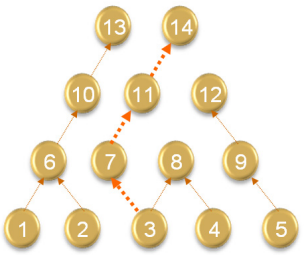
1. <verb.social>S: (v) **conduct**, carry on, deal (direct the course of; manage or control) *"You cannot conduct business like this"*
   1. <verb.social>S: (v) manage, deal, care, handle (be in charge of, act on, or dispose of) *"I can deal with this crew of workers"; "This blender can't handle nuts"; "She managed her parents' affairs after they got too old"*
      1. <verb.social>S: (v) control, command (exercise authoritative control or power over) *"control the budget"; "Command the military forces"*
2. <verb.creation>S: (v) **conduct**, lead, direct (lead, as in the performance of a composition) *"conduct an orchestra: Barenboim conducted the Chicago symphony for years"*
3. <verb.social>S: (v) behave, acquit, bear, deport, **conduct**, comport, carry (behave in a certain manner) *"She carried herself well"; "he bore himself with dignity"; "They conducted themselves well during these difficult times"*
   1. <verb.contact>S: (v) hold, carry, bear (support or hold in a certain manner) *"She holds her head high"; "He carried himself upright"*
   2. <verb.social>S: (v) act, move (perform an action, or work out or perform (an action)) *"think before you act"; "We must move quickly"; "The governor should act on the new energy bill"; "The nanny acted quickly by grabbing the toddler and covering him with a wet towel"*
4. <verb.motion>S: (v) lead, take, direct, **conduct**, guide (take somebody somewhere) *"We lead him to our chief"; "can you take me to the main entrance?"; "He conducted us to the palace"*
5. <verb.motion>S: (v) impart, **conduct**, transmit, convey, carry, channel (transmit or serve as the medium for transmission) *"Sound carries well over water"; "The airwaves carry the sound"; "Many metals conduct heat"*
6. <verb.creation>S: (v) **conduct** (lead musicians in the performance of) *"Bernstein conducted Mahler like no other conductor"; "she cannot conduct modern pieces"*

(그림 1) conduct 동사에 대한 워드넷 매핑

그림 1은 "conduct" 동사가 워드넷에 매핑될 때 발생하는 다중 매핑의 예를 보여준다. "conduct"는 워드넷에서 6개의 동사 신셋에 매핑된다. 또한 각각의 신셋이 상위어로 추상화되어가는 과정에서 3번째 신셋처럼 여러 개의 상위

어를 가지고 있는 경우도 존재한다. 따라서 문장 내의 단어가 가지고 있는 가장 상위 개념의 의미를 결정하기 위해서 신셋 매핑과 상위어 매핑의 의미 애매성 해결이 필요하다.

워드넷 매핑 시에 발생하는 다중 매핑의 문제점을 해결하기 위해서 문장 내에 존재하는 연관관계 주변의 문맥정보를 활용하는 상위어 시퀀스 클러스터링 방법을 사용한다.



(그림 2) 상위어 시퀀스 생성 그래프

그림 2는 동사를 워드넷에 매핑할 때 상위어 시퀀스를 생성하는 과정을 보여주는 그래프이다. 그림 2에서 보이듯이 초기 매핑된 신셋의 최 상위 개념을 찾아 올라가면 여러 갈래의 패스가 존재하는 것을 알 수 있다. 초기 매핑된 신셋(1, 2, 3, 4, 5번 신셋)부터 최상위 신셋(8, 12, 13, 14번 신셋)까지의 연결된 패스를 상위어 시퀀스라고 부르는데, 상위어 시퀀스는 패스가 통과하는 모든 신셋의 관련 정보를 누적해간다. 최종적으로 상위어 시퀀스는 각각의 신셋이 보유하고 있는 유의어 리스트, 정의문, 예제문 등을 자신의 속성 정보로 갖게 된다.

그림 2의 상위어 시퀀스 생성 그래프를 통하여 생성된 상위어 시퀀스 리스트는 표 1과 같다.

<표 1> 상위어 시퀀스 리스트

상위어 시퀀스 아이디	상위어 시퀀스
1	1-6-10-13
2	2-6-10-13
3	**3-7-11-14**
4	3-8
5	4-8
6	5-9-12

문장을 입력받아서 디스크립터에 대한 상위어 시퀀스(이하 주 상위어 시퀀스) 리스트와 문장을 구성하는 나머지 다른 용어들에 대한 상위어 시퀀스(이하 보조 상위어 시퀀스) 리스트를 생성한다. 이렇게 생성된 상위어 시퀀스 리스트를 대상으로 클러스터링을 수행한다. 이때 주 상위어 시퀀스를 초기 클러스터의 중심점(centroid)으로 설정한다. 따라서 클러스터의 개수는 주 상위어 시퀀스의 개수가 된다. 클러스터링을 반복하면서 주 상위어 시퀀스와 보조 상위어 시퀀스가 공통된 개념을 표현하는 인스턴스끼리 서로 모여지게 되고 이것이 곧 클러스터가 된다. 클러스터링이 완료되면 최종적으로 생성된 클러스터를 크기별로 정렬하고 크기가 가장 큰 클러스터를 선택하여 그것의 중심점과 가장 유사한 주 상위어 시퀀스를 해당 동사구의 개념을 대표하는 상위어 시퀀스로 결정한다. 클러스터 선택 및 중심점과 가장 유사한 주 상위어 시퀀스 선택은 해당되는 대상이 나올 때까지 차 상위 순서로 넘어간다. 상위어 시퀀스가 결정되면 해당 디스크립터의 추상화된 개념인 연관관계가 확정된다. 그림 2와 표 1은 상위어 시퀀스 아이디 3이 최종적으로 선택된다는 것을 보여준다. 따라서 신셋 아이디 14가 해당 디스크립터의 가장 추상화된 개념으로 선택된다. 상위어 시퀀스의 클러스터링을 이용한 의미 애매성 해결 방법은 애매성이 발생하는 각각의 영역을 따로따로 고려하지 않고 다중 신셋 매핑과 다중 상위어 매핑의 문제를 한 번에 해결할 수 있다. 알고리즘의 의사 코드(pseudo code)는 다음과 같다.

```
void identifySense(String sentence, String verbPhrase) {
  String validVerb[] = extractValidVerb(verbPhrase);
  HypernymSequence[] mainHypernymSequence =
makeHypernymSequence(validVerb);
  if (mainHypernymSequence.length == 0) {
   print("There is no valid sense!");
  }
  else {
   String validToken[] = tokenizeSentence(sentense);
   HypernymSequence[] subHypernymSequence =
makeHypernymSequence(validToken);
   int hSeqId;
   if (subHypernymSequence.length == 0) {
    hSeqId = calculateSentenceSimilarity(sentence,
mainHypernymSequence);
   }
   else {
    hSeqId = clusterHypernymSequence(mainHypernymSequence,
subHypernymSequence);
   }
   String abstractRelation =
mainHypernymSequence[hSeqId].getAbstractRelation();
   print("AbstractRelation=" + abstractRelation);
  }
 }
```

4. 결론 및 향후 연구

지금까지 상위어 시퀀스의 클러스터링을 이용한 단어의 의미 애매성 해결 방안에 대해서 살펴보았다. 이 방법은 상위어 시퀀스를 생성할 때 워드넷의 신셋 정보를 수집하여 특성 정보로 활용하기 때문에 패스를 통과하면서 누적되는 자질 정보의 양이 많을수록 성능이 좋게 나타난다. 하지만 이러한 방법은 상위어 시퀀스의 길이가 상대적으로 짧은 경우에는 자질 정보를 충분하게 추출하지 못하는 단점이 있다. 따라서 향후에는 상위어 시퀀스의 자질 정보를 충분하게 추출하지 못하는 경우에 대비할 수 있는 방법의 연구가 필요하다.

참고문헌

[1] WordNet, http://wordnet.princeton.edu

[2] EuroWordNet, http://www.illc.uva.nl/EuroWordNet

[3] ROBERTO NAVIGLI, "Word Sense Disambiguation: A Survey", ACM Computing Surveys, Vol. 41, No. 2, Article 10, 2009.

[4] C. H. Jeong, S. P. Choi, and Y. S. Choi, "Fundamental Study on Extracting Relations between Technical Terms using Verb-based Patterns", Proceedings of ICKIMICS2009, Vol. 2, No. 1, pp. 74-77, 2009.