어휘지도(UWordMap)를 활용한 명사와 용언의 다의어 중의성 해소

신준철○, 옥철영
울산대학교, 울산대학교
ducksjc@nate.com, okcy@ulsan.ac.kr

Noun and Verb Polysemy Word Sense Disambiguation Using UWordMap

Joon-Choul Shin○, Cheol-Young Ock
Ulsan University, Ulsan University

요 약

컴퓨터를 이용하여 명사와 용언의 의미를 자동으로 분별하는 것은 기계번역이나 검색 등의 기술에서 아주 중요한 기반 기술이다. 최근에 동형이의어 분별에 대한 연구 결과로 약 96%의 정확률을 보이는 시스템이 개발되었으나, 다의어 분별에 대한 연구는 아직 초기 단계로 일부 어휘만을 한정하여 연구되고 있다. 본 논문에서는 어휘지도를 이용하여 다의어를 분별하는 방법을 연구하였고, 어휘지도에 등록된 모든 일반 명사와 용언을 대상으로 실험하였다. 제안된 알고리즘은 문장에서 나타나는 명사와 용언의 관계를 어휘지도에서 찾고, 그 정보를 기반으로 다의어를 분별하였다. 아직은 그 정확률이 실용적인 수준이라고 볼 수는 없지만, 전체 다의어를 대상으로 실험하였고, 그 실험 결과를 분석함으로써 앞으로의 다의어 분별 연구 방향에 도움될 것으로 판단된다.

주제어: 다의어, 중의성 해소, 태깅, 어휘지도, UWordMap

1. 서론

정보 기술이 발달하면서 검색이나 기계번역 등의 자연어처리 기술의 필요성이 점점 커지고 있다. 이런 자연어처리 기술들이 발전하기 위해서는 형태소 분석이나 의미 중의성 해소 같은 기반기술들이 필요하며 관련하여 많은 연구들이 이미 존재한다. 이 중에서 형태소 분석은 이미 다양한 연구 결과들이 있으며 그 정확률이 매우 높은 편이다. 한국어를 대상으로 한 의미 중의성 해소에 관한 연구는 초기 단계이다.

의미 중의성 해소는 동형이의어(同形異義語)와 다의어(多義語) 두 단위로 구분할 수 있으며, 동형이의어는 형태는 동일하지만 어원이 달라 의미들이 전혀 다른 단어를 뜻하고, 다의어는 하나의 단어에서도 의미가 세분되어 있는 단어를 의미한다. 예를 들어서 '배'는 동형이의어 수준에서 {신체부위, 기계(운송수단), 과일} 등으로 의미가 나뉘어지고 사전에서는 각각 {배_01, 배_02, 배_03}과 같이 어깨번호로 구분한다. '배_01' 다의어에서는 "사람이나 동물의 몸에서 위장, 창자, 콩팥 따위의 내장이 들어 있는 곳으로 가슴과 엉덩이 사이의 부위."라는 의미(배_01①)가 있고, "아이가 드는 여성의 태내(胎內)."와 같은 의미(배_01④) 등이 있다. 즉, 동형이의어 수준이 더 포괄적이고, 다의어 수준은 더 세부적이다.

자연어처리의 많은 영역에서 동형이의어 수준의 중의성 해소로도 만족스러운 결과를 제공할 수 있다. 그러나 일부 영역과 일부 어휘에서는 반드시 다의어까지 정확하게 분별해야만 하는 경우도 있다. 예를 들면 "곤란한 일이 생겼다."라는 문장에서 '일'은 일반적으로 '문젯거리'를 뜻하지만, "요즘은 일이 없어서 가난하다."에서 '일'은 "돈을 버는 활동이나 그 대상"를 의미한다. 이 두 문장을 번역하거나 검색의 대상이 되게 색인을 하려면 '일'의 정확한 의미를 분별하는 것이 중요하다. 그러나 동형이의어 수준에서는 두 문장에서 모두 '일_01'로 동일하고 다의어 수준에서 다르게 구분된다. 따라서 다의어 분별 기술이 반드시 필요하게 된다.

동형이의어에 대한 연구는 이미 진행된 바가 있으며, 약 96%의 높은 정확률을 보이고 있다. 그러나 아직까지 다의어에 대한 연구는 그 수가 매우 적으며, 기존의 한국어 다의어 분별 연구들은 일부 어휘만 한정하여 실험하였다. 본 논문에서는 어휘지도(UWordMap)을 이용하여 다의어를 분별하는 방법을 제안하며, 이 방법의 성능을 실험하기 위하여 모든 일반 명사와 용언을 대상으로 정확률을 측정하였다.

2. 관련 연구

의미 중의성 해소를 위한 연구들은 크게 두 부류의 연구로 나눌 수 있는데, 첫 번째로 대용량 말뭉치를 바탕

으로 베이지안 분류기, 결정 트리, 신경망, CRF, SVM등의 기계학습 기법을 이용하는 연구이다. 두 번째로 기계가독형 사전, 시소러스, 온톨로지, 의미망, 연어 등의 정보를 이용한 지식기반의 기법을 이용하는 연구이다. 기존의 의미 중의성 해소와 관련된 논문의 대부분은 동형이의어 분별에 관한 것으로, 기계학습 기법을 이용하고 있다[1,2,3,4,5].

용언의 의미 중의성 해소에 관한 논문 중 종속격 정보와 공기 빈도를 활용한 연구가 있다[4]. 이 논문은 동형이의어 수준의 분별에 대한 것이며, 12개의 동사를 대상으로 98.7%의 정확률을 보였다. 그리고 문맥에서 추출한 가중치 정보를 이용한 모델을 제안하는 연구가 있었다[5]. 이 연구에서는 약 300만 어절의 사전에서 추출한 공기 관계의 문맥에서 얻은 품사정보와 거리정보를 사용하였다. 이 논문에서는 다의어 수준의 동사(감다, 피우다, 빠지다, 타다) 4개를 대상으로 실험하였으며, 실험결과 84%의 내부실험 정확률과 75%의 외부실험 정확률을 보였다.

최근에는 UWordMap을 이용하여 용언만을 대상으로 다의어 수준의 분별을 시도한 연구가 있었다[6]. 이 연구에서는 문장에서 용언 앞의 논항과 그 명사를 추출한 다음에 용언-명사-논항 정보를 UWordMap에서 찾아서 일치하는 다의어로 태깅하였다. 이 방법은 반드시 용언의 앞에 명사가 위치한 경우에만 적용되었기 때문에 용언이 첫 어절로 나타난 문장은 제외되는 등의 문제가 있었다. 그러나 지식기반 기법을 사용하여 다의어를 분별하였다는 점에서 의미가 있다.

본 논문에서는 UWordMap을 이용하여 용언의 다의어를 분별하면서 동시에 그와 관련된 명사의 다의어도 분별하는 방법을 연구하였다.

3. 한국어 어휘지도 UWordMap

한국어 어휘지도인 UWordMap은 표준국어대사전을 기반으로 한국어 어휘의 다의적 수준의 통사-의미 관계를 네트워크로 연결한 어휘지식베이스이다. UWordMap은 U-WIN[1]을 기반으로 구축되었으며, U-WIN은 품사별로 어휘간의 의미관계(상위어, 하위어, 반의어, 유의어 등)에 중심이 맞춰져 구축되었다. UWordMap은 이러한 U-WIN을 기반으로 구문관계의 어휘들의 의미 제약을 구축한 어휘데이터베이스이다.

UWordMap은 구문관계에 의한 용언과 명사(주격·목적격·부사격) 등의 하위범주화 정보 뿐 아니라, 부사와 용언의 관계, 부사와 부사의 관계, 부사와 명사와 의미제약 관계를 표현하고 있다. 또한 용언의 필수 논항의 의미역이 부여되어 있다. UWordMap은 현재 표준국어대사전의 용례를 기반으로 구축하고 있다.

UWordMap의 구문정보는 용언과 명사어휘망과의 매핑을 통해 구축되었다. 명사어휘망의 명사를 선정할 때는 최소공통상위노드(LCS:least common subsumer)[7]를 선택하였다. 만약 최소공통상위노드의 하위노드 중 용언과 연결될 수 없는 어휘가 포함되어 있을 시에는 이 어휘를 N관계에 포함시켜 해당 용언과의 관계가 연결되지 않도록 설정하였다.

4. 다의어 분별 방법

본 논문에서는 UWordMap의 용언-명사의 하위범주화 정보와 명사의 상위어-하위어 관계의 계층망 정보를 이용하여 다의어 명사와 용언의 의미 중의성 해소 방법을 제안한다. 기본적인 원리는 문장에서 용언이 있다면 그와 관련된 명사를 찾아서 그 둘의 관계를 UWordMap에서 찾고, 그런 관계가 있는 다의어로 태깅하는 것이다.

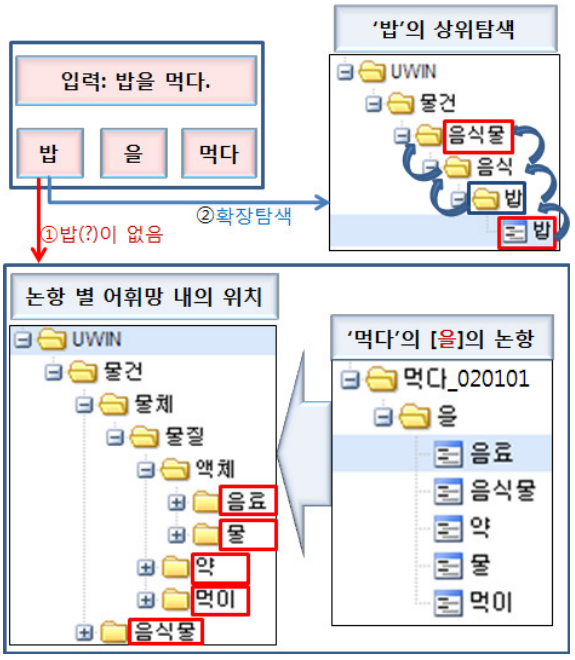

그림 1. 용언의 다의어 분석 예(밥을 먹다)

[그림 1]은 이 방법을 표현한 것으로, 예로 "밥을 먹

1) U-WIN(User Word Intelligent Network)은 한국어의 공통적이고 개별적인 속성을 바탕으로 한국인의 보편적인 인지 체계와 개념 관계를 파악하여 이를 어휘의 의미적·개념적 네트워크를 형성한 온톨로지적 의미망이다. 핵심 구축 대상은 명사, 동사, 형용사이며, 나머지 품사는 부수적인 구축 대상이다.

다”를 들고 있다. UWordMap에서 ‘먹다’는 논항으로 ‘을’을 가지고 있으며 ‘먹다-을’에 해당하는 명사로는 ‘음료’, ‘음식물’ 등이 있다. 이 중에서 ‘음식물’의 하위어에 ‘밥’이 있는 것을 그림에서 확인할 수 있다. ‘밥’에는 다양한 의미가 있지만 ‘음식물’을 상위어로 가지는 ‘밥’이 이 문장에서 문맥적으로 올바르다는 것을 UWordMap을 통해 알 수 있다. 이렇게 ‘밥’의 의미를 정확하게 결정하면서 동시에 ‘먹다’의 의미도 결정할 수 있다. 그림에서는 논항 ‘을’에 ‘음식물’이 있는 ‘먹다_020101’이 문맥적으로 올바르다.

어휘지도를 이용한 상술한 다의어 결정 원리는 기초적인 개념이며, 실제로 그대로 구현하여서는 적용 가능한 문장이 매우 적다. 왜냐하면 실제 문장에서는 “밥 먹었다.” 또는 “밥은 먹었냐”와 같이 조사가 생략되거나 격조사 대신 보조사가 사용되는 경우에는 UWordMap에서 명사-용언의 관계를 찾으면서 논항까지 일치하기를 기대하기는 어렵다. 이 문제를 해결하기 위해서는 논항 정보까지 일치하는 것은 부가적인 자질로 볼 필요가 있다. 이 외에도 ‘음식물’이 최상위 명사에서부터 얼마나 떨어져 있는지에 대한 정보와, ‘음식물’과 ‘밥’ 사이의 거리 등 또한 중요한 정보가 된다. 이런 부가적인 정보들을 모두 적합하게 사용하기 위해서는 하나의 잘 정의된 수학적인 모델을 정의할 필요가 있다.

표 2 자질 함수들

번호	자질 함수
1	해당 동형이의어에서 이 다의어가 첫 번째 의미일 경우에 1을 반환. 그 외에는 0을 반환.
2	하위범주화에서 용언-명사 관계가 있음을 확인하면 1을 반환. 그 외에는 0을 반환.
3	하위범주화에서 용언-논항-명사의 관계가 있음을 확인하면 1을 반환. 그 외에는 0을 반환.
4	하위범주화 관계에 있는 명사의 깊이를 반환 예) ‘음식물’은 최상위 명사까지 거리가 2이므로 2를 반환

다의어 분별은 여러 개의 다의어 중에서 하나를 선택하는 것이며, 어휘가 지닌 다양한 자질값들을 발견하고 이들 자질 간의 가중치를 결정하는 해야 할 것이다. 본 논문에서는 UWordMap을 이용하여 다양한 자질값을 반환하는 자질함수들을 정의하고, 각각의 가중치를 곱하여 최대값을 가지는 다의어로 결정하는 방법을 연구하였다.

<표 2>는 본 논문에서 정의하는 자질 함수들이다. 이 함수들은 0 또는 그 이상의 실수를 반환하며, 모두 UWordMap을 이용하는 것이다. 각 자질에 대한 가중치는 다양한 기존에 알려진 최적화 방법을 사용할 수 있으나, 본 논문에서는 빠른 실험을 위하여 연구자가 직접 실험을 반복하면서 결정하였다.

5. 실험 및 결과 분석

실험을 위하여 최근까지 구축된 UWordMap을 이용하였으며, 형태소를 분석하고 품사와 동형이의어 수준까지의 태깅을 위해서 UTagger[2]를 사용하였다. 정확률 측정을 위하여 표준국어대사전의 용례 말뭉치를 사용하였으며, 형태소 단위로 정답과 오답을 구분하였다. 대상이 되는 형태소는 세종품사 태그가 NNG, VV, VA인 것만을 하였고, 다의어 수준에서 2개 이상의 뜻을 가진 경우에만 대상으로 하였다. 그 결과 대상이 되는 형태소의 수는 약 89만개로 측정되었고, 이 중에서 정확하게 다의어 단위로 분별된 형태소는 약 57만개로 정확률은 65.6%이다. 1번 자질만 사용하였을 때의 정확률은 62.3%로 UWordMap을 사용하여 정확률이 의미가 있는 수준으로 향상되었다는 것을 확인할 수 있다.

표 3 정확률

적용한 자질 함수	정확률(%)
1 (baseline)	62.34
1, 2	65.33
1, 2, 3	65.53
1, 2, 3, 4	65.58

표 4 품사별 정확률

품사	정확률 (%)	정답 형태소	전체 형태소
명사(NNG)	72.93	339,515	465,526
동사(VV)	59.45	200,716	337,606
형용사(VA)	50.48	44,703	88,564

1번 자질을 제외하고 각 자질별로 정확률에 주는 영향을 실험하였으며, 가장 중요한 자질은 2번으로 정확률을 약 3% 향상시키는 것으로 나타났다. 논항까지 일치하는지를 확인하는 3번 자질은 정확률을 약 0.1% 향상시키고, 4번은 약 0.05% 향상시키는 것으로 확인되었다.

품사별로 정확률을 측정하였으며 그 결과라 표 4에 나타나있다. 가장 정확한 품사는 명시이며 약 73%이다. 반

2) UTagger는 울산대학교 한국어처리연구실에서 개발한 품사 및 동형이의어 태깅시스템이다.

면에 형용사는 정확률이 가장 낮았으며, 측정 대상이 되는 형태소의 수도 가장 적었다.

표 1에는 없지만 명사와 그 상위어 사이의 거리 정보를 자질 함수에서 사용하려는 시도를 하였다. 예를 들어서 '가건물을 짓고'에서 '가건물'은 '짓다'와 직접적인 관계는 없으나 그 상위어인 '건물'은 '짓다'와 관계가 있다. 이 때 '가건물'과 '건물' 사이의 거리의 역수를 자질 함수로 사용하고자 하였으나 의미있는 수준의 효과가 나타나지 않았다. 이 부분은 더 심도 있는 연구가 필요할 것으로 보인다.

오답의 유형을 실험자가 일일이 수작업으로 분석하였으며, 대체적으로 문장에서 명사와 관련되는 용언을 알고리즘이 올바르게 찾지 못하거나, 문장에 그런 용언이 없는 경우가 많았다. 예를 들어서 "그에게 10점을 가해도 낙방이야."에서 '낙방'과 관련된 용언은 존재하지 않는다. 따라서 본 논문에 제안하는 방법으로는 이런 경우를 처리할 수 없는 한계를 보였다.

그 다음으로 눈에 띄는 오류 유형으로는 용언-명사 관계를 UWordMap에서 찾을 수 없는 경우다. 예를 들어서 "참기름 따를 때 가에 흘리지 않도록 조심해라."에서 명사 '가'와 용언 '흘리다'의 관계를 UWordMap에서 전혀 찾을 수가 없었다. 심지어 '가'의 상위어 정보를 이용하여도 마찬가지였다. 이것은 UWordMap이 아직 구축중이며 정보가 빈약한 부분이 존재하기 때문이다.

6. 결론

본 논문에서는 UWordMap을 이용하여 명사와 용언의 다의어를 분별하는 연구를 진행하였다. UWordMap의 상위어-하위어 정보와 용언-논항-명사의 하위범주화 정보를 이용하였으며, 유연하게 적용할 수 있도록 다양한 자질 함수들을 정의하였다. 그 결과 대부분의 문장에서 이 방법론을 적용할 수 있었다.

기존의 연구들과 달리 다의어 수준에서 의미가 구분되는 모든 일반 명사와 용언을 대상으로 실험을 하였으며, 그런 이유로 기존 연구 결과에 비하여 정확률 면에서 다소 낮은 편이다. 그러나 모든 일반 명사나 용언에 대해서 다의어 분별을 시도하였다는 것과, 다양한 오류 유형을 분석함으로써 앞으로의 연구 방향을 결정하는 것에 도움이 된다는 점에서 의미가 있는 연구이다.

대표적인 오류 유형들을 분석한 결과, 향후에는 의존관계 정보를 함께 이용하여 다의어 분별의 정확률을 향상시키는 연구가 필요하다. 또한 UWordMap안에서 충분한 정보를 찾을 수 없는 경우를 대비하여 공기어 등의 정보를 이용하는 기존의 방법론을 혼합하는 연구도 필요하다. 특히 등록된 논항과 완전히 일치하지 않더라도 확대 적용할 수 있는 방법을 연구할 필요가 있다.

감사의 글

"이 논문은 2015년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임 (R0101-15-0176)"

참고문헌

[1] 김민호, 권혁철, "한국어 어휘의미망의 의미 관계를 이용한 어의 중의성 해소", 정보과학회논문지,제38권, 제10호, pp.554-564, 2011

[2] 이호, 백대호, 임해창, "분류 정보를 이용한 단어 의미 중의성 해결", 한국정보과학회, 정보과학회논문지(B), 제24권, 제7호, pp.779-789, 1997

[3] 허정, 옥철영, "사전의 뜻풀이말에서 추출한 의미정보에 기반한 동형이의어 중의성 해결 시스템", 한국정보과학회논문지, 제28권, 제9호, pp.688-698, 2001

[4] 박요셉, 신준철, 옥철영, 박혁로, "종속격 정보를 적용한 동사 의미 중의성 해소", 정보처리학회논문지 Part(B), 제18권, 제4호, pp.241-248, 2011

[5] 임수종, 박영자, 송만석, "가중치 정보를 이용한 한국어 동사의 의미 중의성 해소", 한국정보과학회 언어공학연구회 학술발표 논문집, pp.425-429, 1998

[6] 배영준, 옥철영, "어휘지도(UwordMap)를 이용한 용언의 다의어 중의성 해소", 한글 및 한국어 정보처리 학술대회, pp. 167-170, 2013.

[7] Resnik, P., "Using information content to evaluate semantic similarity in a taxonomy", Proceedings of the 14th International Joint Conference on Artificial Inteliigence, Montreal, pp.448-453, 1995