언어적 특징을 반영한 한국어 프레임넷 확장 및 개선

김정욱○, 최기선
한국과학기술원, 기계독습연구실
prismriver@kaist.ac.kr, kschoi@kaist.ac.kr

Expansion and Improvement of Korean FrameNet utilizing linguistic features

Jeong-uk Kim○, Key-Sun Choi
Korea Institute of Science and Technology, Machine Reading Lab

요 약

프레임넷 (FrameNet) 프로젝트는 버클리에서 1997년에 처음 제안했으며, 최근에는 다양한 언어적 특징을 반영하여 여러 국가에서 사용되고 있다. 하지만 문장의 프레임을 분석하는 것은 자연언어처리 전문가들이 많은 시간을 들여야 한다. 이 때문에, 한국어 프레임넷을 처음 만들 때는 충분한 훈련을 받은 번역가들이 영어 프레임넷의 문장들과 그 주석 정보들을 직접 번역하는 방법을 사용했다. 결과적으로 상대적으로 적은 비용이 들지만, 여전히 한 문장에 여러 번 등장하는 프레임 정보를 모두 번역하고 에러를 분석해야 했기에 많은 노력이 들어갔다. 본 연구에서는 일본어와 한국어의 언어적 유사성을 사용하여 비교적 적은 비용으로 한국어 프레임넷을 확장하는 방법을 제시한다. 또한 프레임넷에 친숙하지 않은 사용자가 더욱 쉽게 프레임 정보를 활용할 수 있도록 PubAnnotation 기술을 도입하고 "조사"라는 특성을 고려한 Valence pattern 분류를 통해 한국어 공개 프레임넷 사이트를 개선하였다.

주제어: 자연언어처리, 프레임 의미론, 프레임넷, PubAnnotation

1. 서론

자연언어처리 연구의 필요성은 점차 대두하고 있으며, 이는 자연히 학습과 평가에 사용할 많은 양의 데이터셋을 필요로 한다. 이 현상을 해결하기 위해 버클리 대학[1]에서는 문장의 맥락을 분석하여 프레임이라 불리는 의미 단위들로 분류한 데이터셋을 만들었다. Frame Semantics [1] 에 따르면 문장을 서술하는 단어 중 일부는 프레임을 기술하며 같은 문장에 있는 단어 일부에 해당 프레임에 걸맞은 특정한 역할들을 부여한다. 이렇게 분석되어 프레임넷에 공개된 결과는 의미역 결정(Semantic Role Labeling), 기계 번역(Machine Translation), 정보 추출(Information Extraction), 이벤트 인식(Event Recognition) 등 다양한 자연언어처리 응용 기술에 활용될 수 있다.[2]

하지만 특정 언어 기반의 자연언어처리를 위해서는 해당 언어의 특성을 고려한 데이터셋이 필수적이다. 가장 연구가 활발히 진행된 영어 프레임넷은 전문가가 직접 태깅한 170,000개 이상의 문장이 포함되어 있으며 다른 언어권에도 언어적 개성을 고려한 프레임넷이 구축되었다.[5-7] 2014년에 구축된 한국어 프레임넷 초기 버전의 경우, 훈련된 전문가에 의해 영어 프레임넷 문장이 한국어 프레임넷 문장으로 직접 번역되는 형태로 구축되었다.[3] 이 방법은 임의의 문장에서 프레임 정보를 추출하는 것보다 상대적으로 적은 비용이 소요되지만, 여전히 프레임 정보가 손실되지 않았는지를 확인하는 재검증 작업이 필요하다.

본 연구에서는 언어적 유사성이 존재하는 서로 다른 두 개의 언어 사이의 프레임넷 자원을 확장하는 방법을 제시하며, 실제 일본어 프레임넷 구조를 활용해 한국어 프레임넷을 확장/구축하는 사례를 설명한다. 또한, PubAnnotation[8]을 도입하여 한국어 프레임넷 홈페이지 인터페이스를 개선하고자 한다.

2. 관련 연구

최초로 구축된 프레임넷인 영어 프레임넷을 비롯하여 독일어, 일본어 등 대부분의 프레임넷은 해당 언어에서 활용할 문장들을 뽑은 후, 전문가들이 직접 프레임 정보를 입력하는 방법으로 작성되었다.[1,4]

스웨덴어 프레임넷처럼 수동 구축이 아닌 반자동적 방법을 사용하여 프레임넷을 개발한 예시 또한 존재한다.[5] Toneli와 Pianta는 프레임넷의 의미정보를 WordNet에 대응시키는 방법으로 프레임 정보를 영어에서 이탈리어로 옮기는 사영 알고리즘을 개발하였다.[6] Hartmann은 Wiktionary 의 translation / disambiguation 정보를 활용하여 서로 다른 두 언어간의 프레임 정보를 Wikctionary를 거치는 방식으로 전달하는 방법을 제안하였다.[7]

한국어의 경우에는 영어 프레임넷의 4,025개 문장을 추출하여 번역가들을 통해 프레임을 포함한 문장 정보를 전부 수동 번역하는 방식으로 최초의 한국어 프레임넷을 구축하였고 홈페이지를 통해 이를 공개함으로써 한국어 자연언어처리 연구에 활용할 기틀이 되었다.[3]

1) https://framenet.icsi.berkeley.edu/

3. 한국어 프레임넷 현황

기존의 한국어 프레임넷은 영어 프레임넷의 프레임 정보를 보유한 문장들의 카테고리 중에서 임의로 선택된 총 4,025개의 한국어 문장을 보유하고 있으며, 이 문장들은 전문가에 의해 상세하게 서술된 가이드라인에 따라 번역되었다.[3] 명시된 가이드라인에 따르면 모든 프레임 정보는 문장의 어순이 바뀌더라도 손실되거나 의미에 변형이 가해지지 않으면서, 전체 문장의 의미도 보존되어야 한다. 이렇게 번역된 문장들은 여러 번의 전문가 평가를 거친 후에 한국어 자연언어처리에 활용될 수 있도록 정제되어 한국어 프레임넷 홈페이지[2]에서 공개되었다.

프레임 정보가 담긴 문장과 더불어 프레임넷에는 프레임 인덱스(frame index)라 불리는 요소가 필요하다. 프레임 인덱스란 같은 의미를 가지는 프레임을 묶어서 정의하고 및 필수적인 구성 요소 (core frame element) 와 선택적으로 나타나는 요소 (non-core frame element) 들을 나열하는 것이다. 이러한 프레임 요소(frame element)는 해당 프레임에서 정해진 역할을 가지는 하나의 단어 또는 연속한 여러 개의 단어를 뜻한다. 현재 한국어 프레임넷은 영어 프레임넷을 번역하는 방식으로 구축되었기 때문에 영어 프레임넷의 프레임 인덱스 정보를 그대로 사용하고 있다. "출석" 을 나타내는 프레임 인덱스인 "Attending" 에 대한 정의 페이지는 아래 그림 1에서 살펴볼 수 있다. 해당 프레임 인덱스에 따르면, 출석 (Attending) 이라는 프레임에는 출석하는 대상(Agent)와 출석하고자 하는 사건(Event)가 주요 프레임 요소이고, 선택적으로 "circumstances" 나 "duration" 이 올 수 있음을 보여주고 있다.

그림 1 - 영어 프레임넷의 프레임 인덱스

프레임 정보를 지닌 문장, 프레임 인덱스 목록 이외에도 프레임넷에서는 Lexical Unit(LU) 정보를 추가로 제공한다. LU란 특정 프레임을 유도하는 프레임 요소를 의미한다. 예를 들어, "The company's profit falls 10% due to the accident." 같은 문장에서 "fall" 은 "change_position_on_a_scale", 즉 어떤 수치가 하락한 현상에 해당하는 프레임의 LU가 될 수 있다. 이처럼 한국어 프레임넷에 존재하는 모든 문장에 등장하는 LU들의 원형 총 7,130개는 가나다 순으로 분류되어 리스트의 형태로 나열하였다. 초기 버전의 한국어 프레임넷은 버클리 대학의 영어 프레임넷 디자인을 그대로 사용하였으며, 같은 역할을 하는 영어 프레임넷의 LU 인덱스[3]는 아래 그림 2와 같다. 한국어 LU 인덱스 또한 마찬가지로 검색 및 자음을 통한 색인을 제공한다.

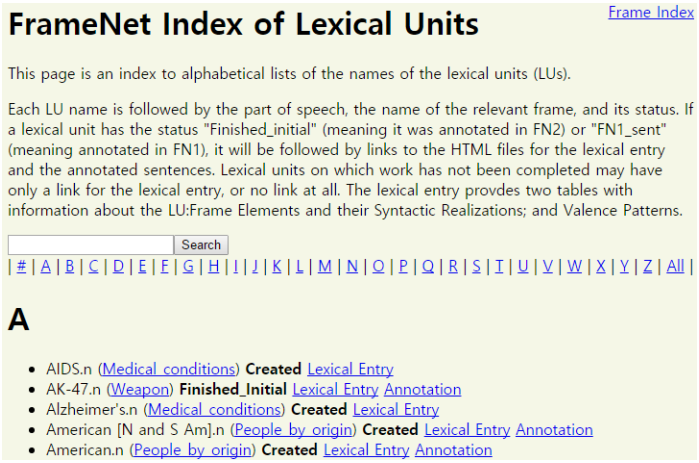

그림 2 - 영어 프레임넷의 Lexical Unit (LU) Index

4. 일본어 프레임넷을 통한 한국어 프레임넷 확장

4.1. 일본어 프레임넷 데이터에서 프레임 정보 추출

일본어 프레임넷[4]에서는 HTML 형태로 색깔과 첨자를 사용해 프레임 정보를 표현한다. 실제 일본어 프레임넷에 존재하는 문장 "梅雨は すでに 明け 、 九州 地方 は 一気に 夏 模様 で ある 。" "장마는 이미 끝나, 규슈 지방은 단숨에 여름이 올 듯한 모양이다." 라는 문장을 예시로 들어 한국어 프레임넷의 문장으로 변환하는 모습을 보이고자 한다. 위 문장의 4개 프레임 정보는 아래의 그림 3에서 확인할 수 있다.

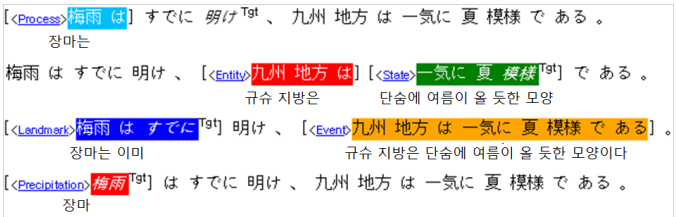

그림 3 - 일본어 문장의 프레임 정보

2) http://framenet.kaist.ac.kr/

3) https://framenet.icsi.berkeley.edu/fndrupal/index.php?q=luIndex

작업의 첫 단계로 위와 같은 HTML 파일을 분석하여, 각 프레임 요소의 위치와 이름을 파악한다. 그리고 프레임마다 프레임 요소들의 시작 위치와 끝 위치를 기준으로 가상의 경계 (boundary)를 세워 원래의 문장을 여러 부분으로 나누는 작업을 진행한다. 서로 다른 프레임에서는 경계가 같은 위치에 나타날 수도 있다. 주어진 문장에 대해 이 과정이 이루어진 모습은 그림 4에서 확인할 수 있다. 파싱 결과로 얻어졌던 다른 프레임 요소 정보는 복구할 수 있도록 따로 보관한다. 예를 들어, 첫 프레임에 있는 "process"에 해당하는 프레임 요소인 '梅雨は'가 첫 번째 경계와 두 번째 경계 사이에 있음을, 그리고 이탤릭체로 표현된 "target"에 해당하는 프레임 요소인 '明け'가 세 번째 경계와 네 번째 경계사이에 있다는 정보가 저장된다.

|梅雨 は| すでに| 明け| 、 九州 地方 は 一気に 夏 模様 で ある 。
梅雨 は すでに 明け 、| 九州 地方 は| 一気に 夏| 模様| で ある 。
|梅雨 は| すでに| 明け 、| 九州 地方 は 一気に 夏 模様 で ある| 。
|梅雨| は すでに 明け 、 九州 地方 は 一気に 夏 模様 で ある 。

그림 4 - 프레임을 기준으로 나뉜 원 문장

4.2. 경계로 나뉜 문장 번역

4.1에서 프레임 요소의 양 끝마다 형성된 경계들을 한 문장에 모아 전문 번역가에게 번역을 의뢰했다. 이 과정에서도 같은 위치에 있는 경계는 하나로 통합하고, 기존의 프레임 요소들의 위치를 경계 인덱스의 형태로 변환한다. 그림 4의 프레임 정보가 담긴 문장들을 통합한 그림 5의 문장에서, "process"에 해당하는 단어 "梅雨は" 는 첫 번째 경계에서 시작하여, 세 번째 경계까지 사이에 있다는 식으로 위치 정보를 보관하는 것이다.

|梅雨| は| すでに| 明け| 、| 九州 地方 は| 一気に 夏| 模様| で ある| 。

그림 5 - 번역할 문장 예시

전문 번역가에 의해 그림 5와 같은 형태의 문장이 원문의 의미를 그대로 가지며, 기존의 문장과 같은 경계 개수를 보유하고 있는 형태인 그림 6으로 번역된다. 특히 각 경계 사이에 존재하는 단어구들은 임의로 지워지거나, 추가되거나, 위치가 바뀌지 않고 상대적 경계 인덱스를 유지하며 한글로 번역이 실행된다.

|장마| 는| 이미| 끝나|,| 규슈 지방 은| 단숨에 여름| 이 올 듯한 모양| 이다|.

그림 6 - 번역된 문장 예시

이러한 기준이 가능한 이유는 한국어와 일본어가 비교적 어순에 자유로운 언어이며, 단어들의 상대적 위치보다는 단어에 덧붙여진 조사가 기존 단어구의 문장 성분을 결정하는 데에 훨씬 커다란 역할을 하기 때문이다. 예를 들어, 일본어의 'は'와 한국어의 '는'은 앞에 있는 단어가 현재 문장의 주어 역할을 하고 있다는 것을 의미한다.

4.3. 프레임 정보가 담긴 한국어 문장으로 변환

그림 6에서처럼 번역이 완료된 경계들로 나눠진 문장을 다시 프레임 정보들로 변환하기 위해 위의 과정을 거꾸로 밟아나가야 한다. 변환 과정에서 저장했던 원래의 일본어 문장에 포함된 프레임 요소의 경계 인덱스를 통해 그 프레임 요소가 한국어로 번역된 문장의 어떤 부분인지 찾을 수 있다. 그리고는 각 프레임 요소가 프레임에서 어떤 역할을 하는지에 대해 저장한 내용을 토대로 한국어 프레임 정보를 지닌 문장을 그림 7에서처럼 재구성한다.

[<Process>장마 는] 이미 끝나Tgt , 규슈 지방 은 단숨에 여름 이 올 듯한 모양 이다 .
장마 는 이미 끝나 , [<Entity>규슈 지방 은] [<State>단숨에 여름 이 올 듯한 모양Tgt] 이다 .
[<Landmark>장마 는 이미Tgt] 끝나 , [<Event>규슈 지방 은 단숨에 여름 이 올 듯한 모양 이다] .
[<Precipitation>장마Tgt] 는 이미 끝나 , 규슈 지방 은 단숨에 여름 이 올 듯한 모양 이다 .

그림 7 - 재구성된 한국어 프레임 정보

4.4. 반례 확인

전문 번역가 측에서 번역 검증을 수행하였으므로 영어 문장의 한국어 번역 결과 자체는 충분히 신뢰할 수 있다. 또한 4.2에서 언급한 번역 가이드라인을 전달하면서, 가이드라인을 지키면서 문장을 자연스럽게 번역할 수 없는 경우가 있다면 알려달라고 요청했었다. 하지만 목표로 삼은 1,795개의 문장을 전부 번역하는 동안 그러한 경우는 없었으며, 본 논문에서 주장하는 한국어와 일본어의 언어적 유사성을 활용한 방법론은 신뢰할 수 있다.

5. 프레임넷 웹사이트 인터페이스 개선

5.1. PubAnnotation 도입

기존의 한국어 프레임넷을 공개하는 창구인 웹사이트의 개선도 동시에 이루어졌다. 초기 웹사이트는 영어 프레임넷의 영향을 받아 프레임 요소 정보를 프레임 인덱스에서 정의된 색깔로 표현하는 영어 프레임넷 방식 그대로를 사용했다. 이로 인해 임의의 문장에서 해당 프레임 요소가 어떤 역할을 가지고 있는지 알아보려면 프레임 인덱스 파일을 참조해야 한다. 그림 8에 나타난 영어 프레임넷의 예시에서 '22%'에 해당하는 프레임 요소가 무엇인지 확인하기 위해서는 검정색으로 표기된 LU인 "fell"에 해당하는 프레임 인덱스의 색깔 표를 확인해야 한다. 실제로 그림 9에 나타낸 "fell"의 프레임 인덱스 표를 확인해보면 해당 프레임 요소는 core frame element 인 "Difference"에 해당한다는 것을 알 수 있다.

기존 색깔을 통해 프레임 요소를 표현하는 방식의 번거로운 작업을 줄이기 위해 한국어 프레임넷에서는

4) http://sato.fm.senshu-u.ac.jp/frameSQL/jfn23/notes/index2.html

[X] Aetna Life and Casualty Co. 's third - quarter net income FELL 22 % to $ 182.6 million , or $ 1.63 a share , reflecting the damages from Hurricane Hugo and lower results for some of the company 's major divisions .

그림 8 - 영어 프레임넷의 프레임 정보 표현

Frame Element	Core Type
Attribute	Core
Circumstances	Extra-Thematic
Containing_event	Extra-Thematic
Correlated_variable	Extra-Thematic
Degree	Peripheral
Difference	Core

그림 9 - fall 에 대한 lexical unit table 일부

PubAnnotation으로 프레임 정보를 나타내는 방식을 도입하였다. Kim and Wang은 2012년에 오픈소스 주석 툴인 PubAnnotation [8]을 공개했다. PubAnnotation은 널리 사용되는 데이터 타입인 JSON을 사용해, 웹 상에서 주어진 문장 일부분에 주석을 입히거나 주석들 사이의 관계를 화살표로 이어주는 기능을 제공한다.

그림 10에서는 그림 8의 문장이 한국어 프레임넷 웹사이트에서 어떻게 표현되었는지를 확인할 수 있다. 화살표의 시작점을 확인함으로써 '하락하' 가 "fall" 에 대응되는 현재 프레임의 LU임을 알 수 있다. 그 뿐만 아니라, 현재 프레임의 프레임 요소위치와 그 각각의 역할이 무엇인지를 바로 위에 위치한 주석을 통해 한눈에 확인할 수 있다는 장점을 지닌다.

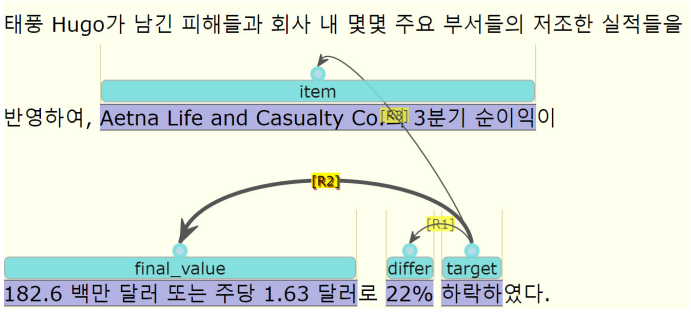

그림 10 - PubAnnotation 활용 예시

XML 파일의 형태로 데이터를 저장하는 영어 프레임넷이나, HTML의 형태로 제공하는 일본어 프레임넷과 달리, 한국어 프레임넷에서는 PubAnnotation으로 변환하기 전의 JSON 형태로 데이터를 제공하고 있다. 이 때문에, 자연언어처리에 프레임넷 자원을 이용하고자 하는 사람들은 XML이나 HTML파싱 대신 주어진 포맷에 따라 간단하게 활용할 수 있다.

5.2. Valence pattern 분류

영어 프레임넷에서는 그림 2에서처럼 LU 인덱스에서 모든 LU의 리스트를 만들어 보관하고 있다. 각 LU에 접근하면 해당 LU의 프레임 인덱스, 간단한 정의, 프레임 요소 표, 그리고 해당 LU가 사용된 문장의 Valence pattern을 확인 및 각 패턴에 해당하는 문장에 접근할 수 있다. Valence pattern 이란 그림 11에서 볼 수 있는 것처럼, LU가 사용된 문장의 프레임 요소를 문장의 상대적 위치순으로 나열해 분류한 것이다. 그리고 세부 항목으로 각 프레임 요소의 품사에 따라 이차 분류가 이루어져 원하는 패턴에 해당하는 실제 문장에도 접근할 수 있다.

Number Annotated	Patterns			
1 TOTAL	Type	Type	Use	Weapon
(1)	DEN --	NP Dep	NP Dep	NP Dep
1 TOTAL	Type	Use	Weapon	
(1)	DEN --	NP Dep	NP Dep	
2 TOTAL	Type	Weapon		
(1)	DEN --	NP Dep		
(1)	N Dep	DEN --		

그림 11 - 영어 프레임넷의 valence pattern

한국어 프레임넷 웹사이트에서는 그림 11과 같이 LU의 원형 및 품사를 표기하는 것과 더불어 해당 LU가 번역되기 전의 LU 형태 정보를 함께 제공하고 있다. 그리고 Valence pattern 표현에 [프레임 요소 이름 / 해당 프레임 요소의 조사]의 형태로 배치하여 한국어의 특징인 조사에 따라 문장에서의 역할이 크게 바뀌는 현상을 반영하고 있다. 그 외에도 각 패턴에 해당하는 문장 리스트에 접근할 수 있는 링크를 표시하고, LU 정보 자체를 저장할 수 있도록 하여 자연어 처리 연구자들이 기존보다 쉽게 활용할 수 있다.

해체되 (VV)

dismantled

01. : [undergoer/JX] + **해체되** (2) Fulltext
02. : [undergoer/JX] + [time/JKB] + [time/JKB] + [manner/JKB] + **해체되** (2) Fulltext

그림 12 - 한국어 프레임넷의 LU 예시

6. 결론

본 논문에서는 일본어 프레임넷의 자료를 활용하여 한국어 프레임넷을 확장한 방법과 한국어 프레임넷 웹사이트의 인터페이스 개선에 대해 논하였다. 문장에 직접 프레임 정보를 입력하는 기존의 방법에 비해, 한국어와 일

본어의 언어적 유사성을 통해 반자동화된 작업 흐름을 따름으로써 비교적 적은 노력으로 한국어 프레임넷 확장에 성공하였다. 또한 다른 언어권 사이에서도 언어적 유사성에 주목한다면 비슷한 접근 방법이 가능할 것이라 기대된다. 또한 프레임넷 웹사이트에 PubAnnotation 을 도입하고 조사를 반영한 Valence pattern 표기를 도입하여 전문 자연언어처리 연구자들은 물론 전문적 지식이 없는 일반 사용자들에게도 보다 나은 사용 경험을 제공한다.

현재 한국어 프레임넷의 가장 큰 해결 과제는 절대적인 데이터의 규모가 크지 않다는 점이다. 비용적 측면에서 영어 프레임넷처럼 직접 프레임 정보를 추가하는 것은 어렵겠지만, 본 논문과 유사한 연구를 통해 한국어 프레임넷을 확장하기 위한 보다 쉬운 방법을 제안하고, 계속 확장해 나간다면 한국어 자연언어처리 연구자들이 다양한 분야에서 자유롭게 사용할 수 있는 언어자원이 되리라 기대한다.

사사

이 논문은 2016년도 정부(미래창조과학부)의 재원으로 정보통신기술진흥센터의 지원을 받아 수행된 연구임. (No. R0101-16-0054, WiseKB: 빅데이터 이해 기반 자가학습형 지식베이스 및 추론 기술 개발)

이 논문은 2016년도 미래창조과학부의 재원으로 한국연구재단 바이오의료기술개발사업의 지원을 받아 수행된 연구임.

참고문헌

[1] Charles J. Fillmore, "Frame semantics." Linguistics in the morning calm (1982): 111-137.

[2] Baker, Collin F., Charles J. Fillmore, and John B. Lowe. "The berkeley framenet project." Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, 1998.

[3] 남세진, 김영식, 박정열, 함영균, 황도삼, 최기선, 영어 FrameNet의 수동번역을 통한 한국어 FrameNet 구축 개발, 한글 및 한국어 정보처리 학술대회 논문집, 2014.

[4] Ohara, Kyoko Hirose, et al. The Japanese FrameNet project: A preliminary report." Proceedings of pacific association for computational linguistics. 2003.

[5] Heppin, Karin Friberg, and Maria Toporowska Gronostaj, "The Rocky Road towards a Swedish FrameNet-Creating SweFN." LREC. 2012.

[6] Tonelli, Sara, and Emanuele Pianta. "Frame Information Transfer from English to Italian." LREC. 2008.

[7] Hartmann, Silvana, Iryna Gurevych, and Ubiquitous Knowledge Processing Lap. "FrameNet on the Way to Babel: Creating a Bilingual FrameNet Using Wiktionary as Interlingual Connection." ACL (1). 2013.

[8] Kim, Jin-Dong, and Yue Wang. "PubAnnotation: a persistent and sharable corpus and annotation repository." Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. Association for Computational Linguistics, 2012.