한국어-프랑스어 자동번역을 위한 과거시제 선어말어미 '-었'의 처리방안

임승희, 노 란*, 홍문표
성균관대학교 독어독문학과, 프랑스어문학과*
{rusilen21,ruangel,skkhmp}@skku.edu

Past Tense Generation in Korean to French Machine Translation

Seunghee Lim, Ran Noh*, Munpyo Hong
Dept. of German Language and Literature, Dept. of French Language and Literature*
Sungkyunkwan University

요 약

본 연구는 현재 개발 진행 중인 다국어 자동통번역시스템에서 발생하는 한국어 과거시제 선어말어미 '-었'의 생성문제를 다루었다. 한국어 과거시제 선어말 어미는 영어와 독일어의 경우에는 대부분 단순과거형으로 생성될 수 있으나, 프랑스어의 경우에는 복합과거의 형식과 반과거의 형식 중 하나를 선택해야 하는 문제가 발생한다. 본 연구에서는 이러한 문제의 해결을 위해 한-프랑스어 코퍼스 분석을 통해 복합과거와 반과거의 올바른 생성을 위한 네 가지의 자질을 선정하였고, 이에 SVM 알고리즘을 적용한 분류기를 구현하였다. 현재까지의 실험결과는 84.45%의 정확률이며 현재 성능개선을 위한 연구가 계속 진행 중이다.

주제어: 기계 번역, 반과거, 복합과거, 과거시제 선어말어미, 상 (aspect), 기계학습

1. 서론[1)]

프랑스어의 과거 시제는 한국어에 비해 세분화되어 있다. 프랑스어의 시제는 단순 과거 (passé simple), 반과거 (imparfait), 복합과거 (passé composé), 대과거 (plus-que-parfait), 전과거 (passé antérieur)의 5가지로 나뉘어져 있으며 문어체에서는 주로 단순 과거를, 구어체에서는 주로 복합과거와 반과거를 사용한다. 본 연구는 구어체 번역을 대상으로 하므로 이 중 반과거와 복합과거를 중점으로 다루고자 한다. 반과거는 과거의 동작이나 상태가 계속되고 있는 것을 나타내며 완전히 끝나지 않은 사실의 묘사를 할 때 사용되는 시제로 미완료상을 표현한다. 반면, 복합과거는 과거에 완료된 동작을 표현한다[1]. 이러한 과거 시제의 구분은 한국어의 과거 시제와 차이를 보인다.

한국어에서 과거는 주로 '-었' 형태소를 통하여 표현되고 프랑스어와는 달리 과거 사태에서 상을 구별하지 않는다. 한국어에서는 과거 사태가 총체적인 관점에서의 상으로 인식되기 때문이다[2]. 따라서 프랑스어를 한국어로 번역할 시에는 모두 '-었' 형태소를 부착하여 과거 시제를 생성하면 번역에 문제가 없지만, 반대의 경우에는 '-었' 형태소를 반과거와 복합과거 중 어느 시제로 번역해야 하는 지에 관한 문제가 발생한다. 본 연구는 이 문제에 대하여 기계학습 방법을 이용한 해결책을 제시하고자 한다.

2. 관련 연구

Kent & Pitt(1996)는 영어와 독일어, 프랑스어 동사구간의 기계 번역을 위한 자질과 형식적 사건 모델(formal event model)을 기반으로 한 의미론적 방식을 제안하였다[3]. 이 연구에서는 동사구의 직관적인 의미가 자질 시스템으로 부호화되는 방법과 이 시스템이 자질 조사 테이블을 통하여 자동 번역 시스템의 모듈식 구조를 지원하는 방법에 대하여 보여준다.

영어와 프랑스어 간의 번역 시에 동사 시제의 모호성을 해소하기 위해 기계 번역 이전에 특정 자질을 태깅하여 실험한 연구도 존재한다. Grisot & Meyer(2014)는 동사 시제의 모호성을 해결하는 자질로 서사성(narrativity)을 제시하고, 이를 수동적인 방식과 자동적인 방식으로 태깅하여 실험하였다[4]. 수동적인 방식에서는 0.91의 카파 값이 얻어졌고, 자동적인 방식에서는 0.72의 F1 스코어가 얻어졌다.

국내에서 한국어-프랑스어 번역에 관한 연구는 주로 문학 작품을 대상으로 이루어지고 있다. 프랑스어 문어체에서는 과거 시제를 대부분 단순 과거로 쓰기 때문에 본 연구의 대상과는 차이가 있다. 그러나 황원미(2003)에서는 문학 작품이 대상이기는 하나 반과거와 복합과거에 대하여 다루고 있다[5]. 이 논문에서는 두 언어의 과거 시제 체계를 분석하고 프랑스어의 반과거와 복합과거가 한국어로 번역되면서 반복 등장하는 다양한 형태소들의 상적 가치를 용언들의 어휘 자질의 특성에 따라 분류하였다. 그 결과, 진행을 나타내는 보조 용언인 '-고

1) 본 연구는 미래창조과학부 및 한국산업기술평가관리원의 산업융합원천기술개발사업(정보통신)의 일환으로 수행하였음. [10041807, 지식학습 기반의 다국어 확장이 용이한 관광/국제행사 통역률 90%급 자동 통번역 소프트웨어 원천 기술 개발]

있-’, ‘-어 있-’은 프랑스어 반과거 시제의 미완료상을 적절히 표현해주고 있다고 보았다. 그러나 반과거와 복합과거에 공통적으로 등장하는 ‘-었’ 형태소는 문맥적 요소와 함께 고려하여야 한다고 보았다.

3. 자질 선택

본 연구에서는 다른 휴리스틱 규칙을 이용하지 않고 기계학습을 통하여 과거 시제의 처리를 해보고자 한다.

자질은 구축된 한국어와 프랑스어의 병렬 코퍼스를 분석하여 선택하였다. 먼저, 한국어에서 진행을 나타내는 보조 용언인 ‘-고 있-’을 첫 번째 자질(F1)로 삼았다. 두 번째 자질(F2)은 문장 안에 상태형용사의 존재 여부를 보았고 세 번째 자질(F3)로 1항 술어의 사용 여부를 선택하였다. 마지막으로 술어가 인지, 감각동사인지의 여부를 자질(F4)로 삼았다.

4. 실험

실험에서 사용된 코퍼스는 20,000 문장 규모의 ETRI 한국어-프랑스어 이중언어 말뭉치이다. 본 연구에서는 2만개의 문장 중 반과거와 복합과거 문장을 추출하여 1833 문장 (반과거 336 문장, 복합과거 1497 문장)을 대상으로 하였다. 실험은 SVM에 기반한 분류기의 성능을 10-fold 상호교차검증 방식으로 평가하였다.

표 1. 실험 결과

Class	Precision	Recall	F-Measure
복합과거	0.866	0.959	0.910
반과거	0.646	0.336	0.442

실험 결과 84.45%의 분류 정확률 (Accuracy)을 얻을 수 있었다. 그러나 표 1에서 확인할 수 있듯이 복합과거에 비해 반과거의 처리율이 현저히 떨어지는 것을 확인할 수 있다. 복합과거는 정확도 (precision)와 재현율(recall)은 모두 높은 반면, 반과거는 둘 다 낮다. 특히 재현율의 경우 0.336라는 낮은 수치가 얻어졌고, 이는 반과거가 거의 처리되지 않았음을 의미한다. 또한, 각 자질의 효용성을 알아보기 위하여 특정 자질만 반영한 실험도 진행되었다.

표 2. 자질 별 실험 결과(%)

	F1	F2	F3	F4
Accuracy	82.00	81.67	84.12	81.67

각 실험 결과는 모두 80% 이상의 생성 정확률을 보이지만 이는 코퍼스에서 복합과거 문장이 차지하고 있는 비율 때문이다. 즉, 실제로 반과거를 처리할 수 있는 자질은 ‘-고 있-’의 보조 용언을 찾은 첫 번째 자질(F1)과 1항 술어의 사용 여부를 찾은 세 번째 자질(F3)이라고 볼 수 있다. 이 실험 결과를 바탕으로 F1과 F3만 반영한 실험은 네 자질을 모두 사용한 실험과 같은 84.45%의 분류 정확률을 보였다.

5. 결론

본 연구에서는 한국어에서 프랑스어로 번역 시 발생하는 과거시제 선어말어미의 생성문제를 다루었다. 본 연구에서 제안한 방법론은 기계학습에 기반한 방법이다.

실험 결과, 전체 분류 정확률은 높았지만 이는 복합과거의 처리가 대부분 가능했기 때문으로 반과거는 낮은 정확도와 재현율을 보이며 거의 처리되지 않았다. 또한, 네 가지의 자질 중 반과거의 처리에 효용성을 가졌던 자질은 보조 용언 ‘-고 있-’과 문장 내 1항 술어의 사용 여부였다. 반과거의 처리율이 낮은 이유는 프랑스어와는 달리 한국어에서는 과거 사태에 대해 진행상을 인식하지 않기 때문이다. 보조 용언 ‘-고 있-’을 부착하여 진행상을 표현할 수 있지만 이 자질만으로는 높은 처리율을 보이지 않는다. 때문에 세 가지의 자질을 더하였으나 이 중에서 상태형용사의 존재 여부와 인지, 감각동사의 사용 여부를 보았던 자질 두 개는 반과거를 처리해주지 못하였다. 두 자질이 반과거에서 많이 쓰이지만 복합과거에서도 마찬가지로 많이 쓰이고 있기 때문으로 분석된다. 또한, 실험에 사용된 문장에서 복합과거의 비율이 반과거에 비해 높아 반과거 자질에 대한 충분한 학습이 이루어지지 못했을 가능성도 존재한다.

따라서 후속 연구에서는 코퍼스에 대한 보완과 본 연구에서 판별된 두 자질 이외에 한국어의 문장에서 프랑스어의 반과거에 해당되는 진행상을 판단할 수 있는 자질을 고려해봐야 할 것이다.

참고문헌

[1] 문유찬· 노윤채, “프랑스어 문법의 세계”, 어문학사, 1997.

[2] 이재성, “한국어의 시제와 상”, 국학자료원, 2001

[3] K. Stuart and J. Pitt, Feature-based & model-based semantics for English, French and German verb phrases, Language sciences 18.1, pp. 339-362, 1996.

[4] G. Cristina and T. Meyer, Cross-linguistic annotation of narrativity for English/French verb tense disambiguation, 9th Edition of the Language Resources and Evaluation Conference, No. EPFL-CONF-198436, 2014.

[5] 황원미, "프랑스어 과거시제의 한국어 번역에 관한 연구-소설 [이방인] 을 중심으로.", 불어불문학연구, 제56집, 제2호, pp. 929-957, 2003.