영한 기계 번역 품사 집합과 펜트리뱅크 코퍼스 품사 집합간의 품사 대응

이성욱, 이공주*, 서정연
서강대학교 컴퓨터학과 자연어처리 연구실, *(주) 마이크로소프트

Part of Speech Mapping between Tagset of English-Korean Machine Translation and Tagset of Penn Treebank Corpus

Songwook Lee, Kong Joo Lee*, Jungyun Seo
Natural Language Processing Lab., Dept. of Computer Science, Sogang Univ., *Microsoft Co.

요약

펜트리뱅크 코퍼스를 기계 번역에서 품사 태깅의 통계 정보 추출에 이용하기 위해서는 펜트리뱅크 코퍼스의 품사 집합과 기계 번역의 품사 집합의 품사 대응이 필요하다. 본 연구는 기계 번역의 품사 태그 집합과 펜트리뱅크의 48개의 품사 태그를 서로 적절히 대응하여 펜트리뱅크 코퍼스의 통계 정보를 이용하는 품사 태깅 시스템을 구축하는데 발생하는 문제점과 그 해결 방안을 제안한다.

1. 소개

약 400만 단어를 가진 펜트리뱅크 코퍼스는 영어의 품사 태깅에 유용한 통계 정보를 제공해 왔다[5]. 그러나 방대한 양의 유용한 코퍼스임에도 불구하고 펜트리뱅크의 품사 태그가 기계 번역 등의 응용 분야에서 사용하는 태그와 다르기 때문에, 펜트리뱅크 코퍼스를 품사 태깅에 이용하려면 품사 태그의 대응이 필요하다. 본 연구는 11개의 품사만 이용하는 기계 번역[1]의 품사 태그 집합과 펜트리뱅크의 48개의 품사 태그를 서로 적절히 대응하여 펜트리뱅크 코퍼스의 통계 정보를 이용하는 품사 태깅 시스템을 구축하는데 발생하는 문제점과 그 해결 방안을 제안한다. 펜트리뱅크 코퍼스에서 추출한 확률 정보를 사용하기 위해서 기계 번역의 품사 태그를 적절한 펜트리뱅크의 품사로 대응시키는 것이 필요하다. 품사 대응 후 대응된 품사의 모호성을 해결하기 위하여 통계 모델을 사용한다. 잘못된 품사 대응은 잘못된 품사 태깅 결과를 초래할 수 있기 때문에 품사 대응 단계가 매우 중요하다고 할 수 있다.

본 연구에서는 품사 대응 후 통계 모델 적용 방법에서 잘못된 품사 대응으로 발생하는 문제점과 잘못된 품사 대응이 발생하는 원인을 알아보고자 한다. 또한, 잘못된 품사 대응에도 유연하게 작동할 수 있도록, 입력문장에 대한 품사 결정은 그 단어에 해당하는 펜트리뱅크에서 추출한 품사 빈도 정보만 사용하여 품사를 결정한 후, 결정된 펜트리뱅크 품사와 가장 유사한 기계 번역의 품사를 찾음으로써 품사를 결정하는 방법을 제안한다.

2. 서로 다른 품사의 대응

기계 번역의 11개의 품사 체계는 Longman dictionary의 품사 분류를 따른다[1]. (n : 명사, pron : 대명사, v : 동사, adj : 형용사, adv : 부사, prep : 전치사, interj : 감탄사, conj : 접속사, det : 한정사, aux : 조동사, sm : 문장 부호) 기계 번역 품사에서 펜트리뱅크 품사로 대응은 단어의 시제, 수, 인칭 등과 같은 특성 정보를 참조하여 대응하도록 만든 품사 대응표를 이용한다. 표1은 동사에 대한 품사 대응표이며 전체 품사 대응표 중 일부분이다. 품사대응을 위한 각 단어의 특성 정보는 형태소 분석 모듈로부터 주어진다.

그림 1은 "I work all day."의 품사 대응 후, 품사 태깅을 하는 과정을 나타낸다. 그림 1의 기계번역 품사에서 펜트리뱅크 품사로 대응하는 과정을 살펴보자. 먼저 각 단어의 기계 번역 품사를 품사 대응표를 이용하여 가능한 모든 펜트리뱅크 품사

로 대응시킨다.

표 1. 동사의 품사대응표

펜트리뱅크의 품사	기계번역의 품사	특성정보
VB	v	시제: 원형
VBD		시제: 과거
VBG		시제: 진행형
VBN		시제: 과거완료
VBP		시제: 현재, 인칭: 1,2
VBZ		시제: 현재, 인칭: 3, 수: 단수

I	work	all	day
pron → *PRP*	*v* → *VB*, *VBP*	*adj* → *JJ*	*adj* → *JJ*
	n → *NN*	*adv* → *RB*	*n* → *NN*
		det → *DT*,*PDT*	
		pron → *PRP*	

그림 1 기계 번역 품사의 펜트리뱅크 품사로의 대응

그림 1에서 'work'는 품사 v, 현재 시제, 동사 원형의 특성정보가 있어 각각 현재 시제에 해당하는 펜트리뱅크의 VBP와 동사 원형의 VB로 대응되었다. 이렇게 대응된 품사 범위 안에서 품사 태깅을 위한 확률 모델을 적용한다. 결정된 펜트리뱅크 품사를 다시 기계 번역의 품사로 대응시켜 구문 분석단계로 출력한다. 그림 1에서 밑줄 친 품사는 결정된 품사를 뜻한다.

3. 품사 대응 오류

품사 대응에서 발생하는 오류는 품사 태깅 모듈의 성능에 큰 영향을 준다. 품사 대응에서 발생하는 오류는 그 이유를 크게 두 가지로 나누어 볼 수 있다. 첫째, 품사 결정 모델의 전단계인 형태소 분석 단계에서의 오류로 인해, 품사 대응에 사용하는 특성정보가 부족하여 잘못된 펜트리뱅크의 품사로 대응되는 경우이다. 예를 들어, 그림 1에서 'work'의 시제 정보 중 현재 시제 정보가 없다면 기계 번역의 품사 v는 VB(동사원형)로만 대응된다. 둘째, 특정 단어들의 경우 형태소 분석 단계에서 분석된 품사의 종류가 펜트리뱅크에서 관찰된 품사의 종류보다 많아서 발생하는 경우다. 이 경우에, 품사 대응은 올바르게 되지만 문맥상 올바른 품사에 대한 펜트리뱅크 코퍼스에서 관찰된 확률 정보가 없게 된다. 그래서 품사 태깅에서 올바른 품사가 탈락하게 된다. 이러한 오류도 다시 두 가지 경우로 나눌 수 있다. 한 가지는 단어의 서로 다른 쓰임에 대해 동일하게 태깅된 경우이고 또다른 한 가지는 데이터 부족 문제로 발생하는 경우이다. 전자의 경우는 그림 1의 'all'과 같이 펜트리뱅크에서 DT, PDT, RB로만 태깅된 경우를 들 수 있다. 이 경우에 'all'에 대한 기계번역의 품사, adj와 pron을 각각 대응한 JJ와 PRP는 무조건 최소 확률을 가지게 된다. 만약 adj나 pron이 'all'의 올바른 태그인 문장에서는 항상 이러한 태그를 잘라내어 다음 단계인 구문분석에서 잘못된 분석을 하게된다. 이와 동일한 문제점을 갖는 예는 다음 펜트리뱅크 태그로 태깅한 문장을 살펴보면 알 수 있다. 괄호 안은 밑줄 친 단어의 기계번역의 태그를 뜻한다.

a) This/DT Book/NN is/VBZ expensive/JJ ./. (det)
b) This/DT is/VBZ expensive/JJ ./. (pron)

'this'는 형태소 분석 결과에 따라 a)는 DT로 대응하고 b)는 PRP로 대응한다. 각각의 경우 전자는 지시형용사로 쓰인 경우이고 후자는 지시대명사로 사용된 경우이다. 그러나 펜트리뱅크에서는 위의 예에서와 같이 두 가지 경우를 구분하지 않고 모두 DT로 태깅하였다. b)와 같은 문장에서 'this'가 PRP로 대응된 기계 번역의 품사는 항상 상위 2순위에서 탈락하게 되는 문제점이 발생한다. 이 외에도 많은 단어가 서로 다른 경우에 사용되었지만 동일한 태그로 태깅되었다[6]. 이런 경우에 올바른 품사의 탈락을 막으려면 펜트리뱅크 코퍼스에 이런 경우를 구분할 수 있도록 다시 태깅하여 관찰 빈도를 구해야한다. 아니면 품사 대응 규칙을 매번 수정하여 펜트리뱅크 코퍼스에서 관찰되는 태그로 대응을 하여 올바른 품사가 탈락하지 않도록 해야 한다. 그림 1의 예문에서는 'all'의 adj를 DT로 대응시키면 adj를 탈락시키지 않을 수 있다.

이와 같은 품사대응 후 품사 태깅 방법은 하나의 품사에 대하여 가능한 모든 펜트리뱅크의 품사로 대응을 하고 대응된 품사들 중에서만 가장 확률이 높은 품사를 선택하였다. 그렇지만 형태소 분석의 오류와 대응 정보 부족 등에 의해 품사가 잘못 대응된 경우에 정확한 품사 결정이 어렵다. 이 방법은 품사 대응에서 오류가 생겼을 때는 확률 모델을 적용하여 품사 결정을 하는 의미가 없어진다. 왜냐하면 확률 모델에서 bigram확률과 어휘 확률을 사용하는데 품사 대응에서 오류가 발생하면 엉뚱한 품사의 어휘확률과 bigram확률을 사용하게 되므로 전체 확률 모델 자체가 깨어지는 결과를 초래하기 때문이다.

품사 대응의 오류로 확률 모델이 깨어지지 않게 하기 위해서, 어떤 단어의 가능한 품사들이 펜트리뱅크 코퍼스에서 관찰된 품사들과 동일한 확률 분포를 가진다고 가정한다. 형태소 분석 단계에서 넘겨진 단어의 각 품사에 대한 정보를 이용하지 않고 그 단어를 펜트리뱅크 코퍼스에서 관찰된 품사들로 통계적 품사 태깅을 하여 2-Best 결과의 펜트리뱅크 품사 2개를 결정한다. 그 후에 품사 대응표를 이용하여 결정된 2-Best 펜트리뱅크 품사와 가장 유사한 기계 번역의 품사를 선택한다. 이 방법에서 통계 모델은 품사 대응에 전혀 영향을 받지 않는다. 품사 결정을 위한 단어의 여러 품사를 펜트리뱅크 코퍼스에서 발견되는 모든 품사를 사용하여 통계 모델을 적용함으로 통계 모델 자체의 성능을 보장한다. 즉 형태소 분석의 오류나 품사

대응의 오류가 통계적 품사 결정에 영향을 끼치지 않게 함으로써 통계 모델 자체가 깨어지지 않게 한다. 그 후, 펜트리뱅크 코퍼스에서 관찰되는 태그에 해당하는 기계 번역의 태그를 품사 대응표를 이용하여 찾아 준다. 잘못된 품사 대응이 발생하였을 때 품사 대응표를 수정하여 품사 대응의 오류를 줄일 수 있는 것은 품사 대응 후 통계 모델 적용 방법과 동일하다.

I	work	all	day
CD	VB	DT → det	NN → n
LS	VBP → v	PDT	
NN	NN	RB	
NNP			
PRP →pron			

그림 2. 펜트리뱅크 품사의 기계 번역 품사로 대응

그림 2는 "I work all day." 문장의 통계 모델 적용 후 품사 대응 방법을 사용한 예를 보인다.

어휘 'all'에 대해 결정된 품사 DT를 기계 번역의 adj와 det 두 품사로 대응하는 품사 대응 규칙을 품사 대응표에 추가하여 'all'의 adj 품사를 탈락시키지 않게 할 수 있다. 또 형태소 분석 단계의 오류가 발생하여도 품사 결정에 영향을 줄일 수 있다. 예를 들어 그림 2에서 'work'의 특성 정보 중 현재시제 정보가 없다면 기계 번역의 품사 v에서 VBP는 빠지고 VB로 대응된다. 그러나 통계 모델을 독립적으로 사용함으로써 'work'의 품사가 VBP로 결정된다. VBP는 동사이므로 동사류 태그를 가장 비슷한 태그로 가정하고 VB를 찾는다. 대응 오류가 많이 발생하는 유사 품사 집합을 서로 찾을 수 있도록 하면 품사 결정 시스템의 성능을 개선할 수 있다.

4. 실험 및 평가

본 품사 태깅 모듈은 형태소 분석기로부터 문장의 각 단어의 품사와 시제, 수, 인칭 등의 특성 정보를 입력으로 한다. 문장의 각 단어의 기계 번역 품사를 펜트리뱅크 코퍼스에서 추출한 확률 정보를 이용하여 결정한다. 펜트리뱅크 코퍼스에서 추출한 확률 정보를 사용하기 위하여 기계 번역 시스템의 품사 태그를 적절한 펜트리뱅크의 품사 태그로 대응시킨다. 기계 번역 시스템의 품사 태그가 펜트리뱅크의 품사 태그로 올바르게 대응하는 것이 품사 결정의 정확도에 많은 영향을 준다. 품사 대응의 오류는 형태소 분석 모듈의 분석 결과에서 단어에 대한 동사의 시제, 명사의 수, 형용사, 부사의 비교급 등의 특성 정보를 올바르게 분석하지 못하여 발생하고 기계 번역 품사에서 펜트리뱅크 품사로 대응하는 것 자체의 한계에서 발생한다.

펜트리뱅크 코퍼스를 사용하기 위해 품사 대응이 필요한 경우 품사 대응의 오류로 인해 통계적 품사 결정 모델이 깨어지지 않게 하기 위하여 통계 모델을 독립적으로 수행한 후에 가장 유사한 기계 번역의 품사를 찾는 방법을 제안하였다.

Hidden Markov Model을 이용한 품사 태깅의 bigram확률과 단어 확률 계산에 펜트리뱅크에서 추출한 확률 정보를 사용한다[2,3,4]. 1순위 품사는 경로 기반 품사 결정 방법으로 결정하고 2순위 품사는 상태 기반 품사 결정 방법을 사용하여 결정하는 영한 기계 번역 시스템을 위한 품사 태깅 모듈을 설계한다[2,3]. 300문장, 3683단어에 대하여 성능을 평가한 결과 품사 대응 후 통계 모델 적용 방법은 1순위 97.28%, 2순위까지 99.64%였다. 통계 모델 적용 후 품사 대응 방법에 의한 성능은 1순위 97.77%, 2순위까지 99.67%였다. 두 방법의 성능이 비슷하지만 기계 번역 분야와 같은 응용분야에서 사용할 때, 전단계인 형태소 분석의 오류로 인한 품사 태깅의 오류를 줄일 수 있다는 점에서 통계 모델 적용 후 품사 대응 방법이 더 robust하다고 할 수 있다.

대부분의 오류는 복합어와 미등록어에서 발생하였다. 'South Korea', 'New Zealand'등에서와 같이 '형용사+명사'가 새로운 명사를 뜻할 때 여러 단어를 하나로 묶어주는 것이 필요하다. 이는 형태소 분석 단계에서 해결해야 할 과제라고 생각된다.

데이터 부족 문제로 발생하는 대응 오류는 아예 특정 품사에 대한 빈도수가 발견되지 않은 경우이다. 예를 들어 'quarantine'과 같은 단어는 펜트리뱅크 코퍼스에서 VB로 1회 관찰되었다. 그러나 이 단어는 명사로도 사용되는 단어이다. 이 단어는 통계 모델을 사용해도 항상 명사 태그는 선택이 되지 않는다. 이런 저빈도 단어의 가능한 모든 품사에 대한 확률 정보를 얻을 수 있는 방법이 요구된다.

참고문헌

1. 임철수, 이현아, 최명석, 장병규, 이공주, 김길창, "어휘화된 규칙에 기반한 영한 기계번역 시스템," 한국정보과학회 학술발표논문집 제24권 2호, 1997.
2. G. F. Foster. 1991. Statistical Lexical Disambiguation. Master's thesis, McGill Univ. School of Computer Science, Montreal, Canada.
3. J. Allen. 1995. Natural Language Understanding. The Benjamin / Cummings Publishing Company, Inc.
4. J. Kupiec. 1992. Robust part-of-speech tagging using a hidden Markov model. Computer Speech and Language, 6:225-242.
5. M. Mitchell, S. Beatrice, M. Maryann, 1993. Building a Large Annotated Corpus of English : the Penn Treebank. Computational Linguistics.
6. B. Santorini, 1991. Part of Speech Tagging Guidelines for the Penn Treebank Project.