5LHMM기반 영어 형태소 품사 태거의 도메인 적응 방법

권오욱° 김영길
한국전자통신연구원 소프트웨어연구부문 음성언어정보연구부 언어처리연구팀
ohwoog@etri.re.kr, kimyk@etri.re.kr

Domain Adaptation Method for LHMM-based English Part-of-Speech Tagger

Oh-Woog Kwon° Young-Gil Kim
Natural Language Processing Tram, Electronics and Telecommunication Research Institute

많은 언어처리 시스템에서 전처리를 위하여, 형태소 품사 태거를 이용한다. 형태소 품사 태깅 성능은 이를 이용하는 언어처리 시스템에 중요한 역할을 한다. 이러한 태깅 성능 향상을 위하여, Hidden Markov Model(HMM)[2], Lexicalized Hidden Markov Model(LHMM)[5], Conditional Random Field(CRF) 모델[4], Transformation-based Learning[1], Memory-based Learning[3], 결정 트리[6], Maximum Entropy Model[4] 등의 통계적 학습 방법론들이 연구되었다.

자동번역과 같은 언어처리 시스템들은 최근에 도메인에 적응하여 분석하여 언어처리의 애매성을 줄여 높은 성능을 보이고 있다[7]. 이와 같이 특정 도메인에 적응할 경우에 형태소 품사 태깅과 같은 애매성을 줄여 그 성능을 향상할 수 있다. 하지만, 특정 도메인에 적합한 품사 태깅된 코퍼스의 부재가 문제가 된다. 일반적으로 품사 태깅된 코퍼스는 각 언어권별로 한정된 자원이므로, 코퍼스에 기반한 형태소 품사 태깅 모델들을 학습하지 않은 특정 도메인에 적합하도록 변경하기는 어렵다.

본 논문에서는 특정 도메인에 적합한 품사 태깅된 코퍼스가 없이, 일반적인 품사 태깅된 코퍼스로부터 학습한 형태소 품사 태깅 모델을 도메인에 적응하는 방법을 제안한다. 제안하는 방법은 기존에 학습된 LHMM의 전이확률(transition probability) $P(t_i|t_{i-1},t_{i-2})$와 출력확률(output probability) $(w_i|t_i)$을 특정 도메인에 적합하도록 변경하기 위하여, 특정 도메인의 원시코퍼스를 기학습된 LHMM으로 자동 태깅한다. 그리고, 자동 태깅된 도메인 코퍼스에서 전이확률 $P'(t_i|t_{i-1},t_{i-2})$과 출력확률 $P'(w_i|t_i)$을 추출하여, 기존 학습코퍼스로부터 추출한 확률값들과 자동태깅된 도메인 코퍼스로부터 추출한 확률값들 간의 차이가 큰 어휘와 trigram을 추출한다. 추출된 어휘와 trigram에 관련된 전이확률과 출력확률을 특정 도메인에 맞도록 영어 전문가가 전문가의 언어적 직관으로 휴리스틱하게 조정한다.

LHMM의 출력확률을 조정할 어휘들을 추출하기 위하여, $\mathrm{abs}(P(w_i|t_i)-P'(w_i|t_i))$ 가 임의의 임계치 Wt를 넘는 단어 w_i 들을 추출한다. 출력확률 $P(w_i|t_i)$은 임의의 품사 t_i에서 임의의 단어 w_i가 출력 또는 발생할 확률이다. 예를 들어, 출력확률 $P(w_i$="write"$|t_i$=VB)는 VB(동사원형) 품사일 경우에 "write"이 나타날 확률을 의미한다. 이것은 전문가에 의해서 언어 직관력으로 그 확률을 조정한다고 하더라도 확률 값을 조정하기가 매우 어렵다. 하지만, 반대로 write 단어가 NN(명사)와 VB(동사원형), VBP(현재 복수형 동사)을 품사로 가질 때, "write"이 특정 도메인에서 어떤 품사 분포로 나타날 지에 대해서는 자신만의 언어직관력으로 표현할 수 있다. 즉, 사람의 언어적 직관에 의해서 출력확률 $P(w_i|t_i)$은 표현하기는 어렵지만, 임의의 단어에 대한 품사가 될 확률 $P(t_i|w_i)$은 표현할 수 있다. 이러한 언어적 직관에 의한 영어 전문가에 의한 수정이 가능하기 위하여, 본 논문에서는 출력확률 $P(w_i|t_i)$을 어휘품사확률 $P(t_i|w_i)$을 이용하여 수식 (1)과 같이 표현한다.

$$P(w_i|t_i) = P(t_i|w_i)\times P(w_i)/P(t_i) = P(t_i|w_i)\times f(w_i)/f(t_i) \qquad (1)$$

수식 (1)에서 $f(w_i)$는 단어 w_i가 전체 코퍼스에서 나타나는 빈도수를 의미하고 $f(t_i)$는 품사 t_i의 전체 빈도수를 의미한다. 본 시스템에서는 각 단어와 각 품사에 대한 빈도수 $f(w_i)$와 $f(t_i)$를 저장하고, 또한 어휘품사확률 $P(t_i|w_i)$을 저장하였다가, LHMM에서 출력확률 $P(w_i|t_i)$이 필요한 경우에 수식 (2)에 의해서 계산한다.

전문가에게 추출된 어휘 w_i와 그 어휘가 가질 수 있는 모든 품사들에 대한 $P(t_i|w_i)$를 제공하고, 더불어 자동태깅된 도메인 코퍼스에서 추출한 어휘 w_i에 대한 모든 품사들의 $P'(t_i|w_i)$를 같이 제공한다. 또한, 어휘 w_i에 대한 각 품사로 태깅된 예문들을 도메인 코퍼스에서 추출하여 최대 10문장씩 제공한다. 전문가는 어휘 w_i가 예문들에 제공된 문장에서 그 단어가 어떠한 품사로 태깅되어야 하는가를 표시한다. 임의 단어 w_i가 m개의 품사 $t_1, \cdots, t_m$을 가진다고 할 때, 전문가가 품사 당 10문장씩 올바르게 태깅한 품사에 의하여 수식 (2)과 같이 조정 가능한 어휘품사확률 $P''(t_p|w_i)$을 전문가에 제시한다.

$$P''(t_p \mid w_i) = \sum_{j=1}^{m} (P(t_j \mid w_i) \times f(t_p \mid t_j) / N_j) \quad (2)$$

수식 (2)에서 $f(t_p \mid t_j)$는 품사 t_j으로 태깅된 예문에서 실제 정답이 품사 t_p인 문장 수이며, N_j는 실제 품사 t_p로 태깅된 전체 예문 수이다. 수식 (2)에 의해서, 실제 도메인 코퍼스에서 자동태깅한 예문이 모두 맞으면 기존 학습된 어휘품사확률을 조정하지 않아도 된다. 많이 틀리는 품사에 대해서는 어휘품사확률 조정이 크게 발생한다. 전문가는 수식 (2)에서 제시된 어휘품사확률 $P''(t_p \mid w_i)$을 근거로 하여 기존 어휘품사확률 $P'(t_p \mid w_i)$과 도메인 코퍼스에서 추출한 어휘품사확률 $P'(t_p \mid w_i)$을 총체적을 최종 도메인에 적합한 어휘품사확률을 정한다.

LHMM의 전이확률은 대상언어에 대한 문법적 특성을 나타내고 있다. 본 논문에서는 도메인이 달라지더라도 대상언어가 가지는 기본적 문법적 특성을 그대로 보유하면서, 자주 나타나지 않은 문법적 특성들이 도메인에 따라 나타날 것으로 생각한다. 본 논문에서는 LHMM의 전이확률을 도메인 적응하기 위하여, 먼저 기존 학습코퍼스에서 임의의 임계치 Tt이하의 확률값을 가진 $P(t_i \mid t_{i-1}, t_{i-2})$(확률값 0도 포함)가 도메인 코퍼스에서의 전이확률 $P'(t_i \mid t_{i-1}, t_{i-2})$가 높은 순서로 trigram을 정렬하여, 전문가에 의해 임의의 순서 이상을 대상으로 하여 기존 학습된 전이확률에 추가하여 사용한다. 선택된 trigram은 전문가에 의해 문법적으로 정확한 것인가를 예문을 통하여 확인하여 언어적으로 잘못된 trigram은 배제한다. 선택된 trigram t_{i-2}, t_{i-1}, t_i에 대한 전이확률은 먼저 선택되지 않은 trigram 중에서 전이확률 $P'(t_i \mid t_{i-1}, t_{i-2})$와 가장 가까운 전이확률 $P'(t_q \mid t_{q-1}, t_{q-2})$값을 가진 trigram t_{q-2}, t_{q-1}, t_q을 찾는다. 그리고, 도메인 적응된 전이확률 전이확률 $P'(t_i \mid t_{i-1}, t_{i-2})$은 수식 (3)와 같이 계산하였다.

$$P(t_i \mid t_{i-1}, t_{i-2}) = P'(t_i \mid t_{i-1}, t_{i-2}) \times \frac{P(t_q \mid t_{q-1}, t_{q-2})}{P'(t_q \mid t_{q-1}, t_{q-2})} \quad (3)$$

본 논문에서는 제안한 LHMM 기반의 영어 형태소 품사 태거에 대한 도메인 적응 방법을 실험하기 위하여, 영어 특허문서 도메인에 LHMM 기반 영어 형태소 품사 태거를 적응하는 실험을 하였다.

다음 2가지 시스템에 대하여 비교 실험하였다.

- LHMM 태거: PennTree Bank에서 학습한 LHMM기반의 영어 형태소 품사 태거
- LHMM 도메인 태거: LHMM 태거의 학습된 값을 10만 특허문서 원시코퍼스를 대상으로 하여 6,000 어휘에 대한 어휘품사확률을 조정하고 1,500 trigram에 대한 전이확률을 조정한 도메인 적응 LHMM 기반 영어 형태소 품사 태거 (전문가 3인에 의해서 1개월 작업 분량)

<표 1> 도메인 적응에 대한 비교 실험

비교 대상	단어단위 태깅 정확률	문장단위 태깅 정확률
LHMM 태거	96.98%	49.00%
LHMM 도메인 태거	99.64%	90.50%

본 실험을 위한 실험 집합으로는 100만 특허문서에서 랜덤하게 추출한 200문장(평균 27.48단어)에 대상으로 하였다. 영어 특허도메인에 적응하는 실험을 통하여 단어단위 태깅 정확률 99.64%와 문장단위 태깅 정확률 90.5%의 성능을 보였으며, 도메인 적응하지 않은 형태소 태거보다 단어단위 태깅 정확률 2.66% 향상과 문장단위 태깅 정확률 41.5% 향상을 보였다.

참고문헌

[1] Brill, E., "Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging", Computational Linguistics 21(4): 543-565, 1995.

[2] Brants, T., "TnT - a statistical part-of-speech tagger", Proceedings of the Sixth Applied Natural Language Processing (ANLP-2000), Seattle, WA, pp. 224- 231, 2000.

[3] Daelemans, W., Zavrel, J., Berck, P. and Gillis, S., "MBT: A memory-based part-of-speech tagger generator", Proceedings 4th Workshop on Very Large Corpora, pp. 14-27, 1996.

[4] John Lafferty, Andrew McCallum, and Fernando Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data", Proceedings of the Eighteenth International Conference on Machine Learning 2001, pp. 282-289, 2001.

[5] Ferran Pla and Antonio Molina, "Improving Part-of-speech Tagging Using Lexicalized HMMs". Natural Language Engineering 10(2) 167-189, 2004.

[6] Ma´rquez, L., Padro´, L. and Rodr´ıguez, H, "A machine learning approach to POS tagging", Machine Learning 39(1): 59-91, 2000.

[7] Munpyo Hong, Young-Gil Kim, Chang-Hyun Kim, Seong-Il Yang, Young-Ae Seo, Cheol Ryu, and Sang-Kyu Park, "Customizing a Korean-English MT System for Patent Translation", MT Summit X. 181-187, 2005.