띄어쓰기 비종속 품사 태깅 시스템 개발

모비코앤시스메타(주) 기술연구소
이 경 일 · 안 태 성

Development of POS Tagging System Independent to Word Spacing

Kyung-Il Lee, Tae-Sung Ahn
Technical Research Center, Mobico & Sysmeta, Co., Ltd, Seoul, Korea

요 약

본 논문에서는 입력된 한국어 문자열로부터 형태소를 분석하고, 품사를 태깅하는 방법에 있어 개선된 통계적 모델을 제안하고, 이에 기반한 띄어쓰기 비종속 형태소 분석 및 태깅 시스템의 개발과 성능 평가에 대한 결과를 소개하고 있다. 제안된 통계 기반 품사 태깅 시스템은 입력된 문자열로부터 음절의 띄어쓰기 확률값을 계산하여 유사어절을 생성하고, 유사어절 단위로 사용자 띄어쓰기와 상관없이 형태소 후보 리스트를 생성하며, 인접한 후보 형태소들의 접속 확률 계산에 있어 어절 간 접속 확률과 어절 내 접속 확률을 모두 사용함으로, 최적의 형태소 리스트를 결정하는 모델을 사용하고 있다. 특히, 형태소들의 접속 확률 계산 시 어절 간 접속 확률과 어절 내 접속 확률의 결합 비율이 음절의 띄어쓰기 확률값과 사용자의 띄어쓰기 여부에 따라 자동으로 조절되는 특징을 가지고 있으며, 이를 통해 극단적으로 띄어 쓰거나 붙여 쓴 문장에 대해서도 평균 90% 수준의 품사 태깅 성능을 달성할 수 있었다.

서 론

지난 5년간 인터넷 산업의 성장과 함께 콘텐츠의 생성과 유통이 폭발적으로 증가하였다. 특이할 만한 사항은 이미 존재하던 문서가 온라인을 통해 유통되는 경우 보다는 게시판 등을 통해 인터넷 상에서 실시간으로 저작된 텍스트 콘텐츠의 양이 기하급수적으로 증가하고 있다는 것이다. 이러한 문서들의 경우 상당히 비 문법적이며, 특히 띄어쓰기 및 개행 처리에 있어서 심각한 오류를 포함하고 있다. 이러한 오류는 검색, 자동번역, 텍스트마이닝 등의 언어정보처리에서 가장 큰 걸림돌로 인식되고 있고, 최근에는 대표적 문서 유통 포맷인 PDF 문서의 분석에서도 심각한 문제로 부각되고 있다.

언어정보처리에 있어서 한국어 형태소 분석 및 태깅 기술은 가장 기본적이면서도 중요한 요소로 자리매김하고 있다. 지난 10년간 한국어 형태소 분석과 품사 태깅 부분에 있어서 매우 큰 진전을 보였으며, 현재는 다양한 형태소 분석 및 태깅 방법론 중 대용량 말뭉치에 기반한 통계적 모델이 널리 쓰이고 있다.

통계적인 방법에 기반해 다수 개의 형태소 후보 리스트 중 최적의 형태소 리스트를 선택하는 방법은 대부분 히든 마르코프 모델(hidden Markov's model, HMM)에 기반하고 있으며, 형태소 간의 접속확률은 식(1)과 같이 문맥 확률(contextual probability)과 어휘 확률(lexical probability)의 곱으로 나타낼 수 있다(김재훈, 임철수, 서정연, 1995). 한국어가 영어와 같이 형태소 간 띄어쓰기를 하고 있지 않기 때문에 대부분의 통계 기반 형태소 분석 시스템에서는 형태소의 묶음인 어절에 기반해 HMM을 적용하고 있고, 띄어쓰기 오류를 포함한 문서에 대해서는 매우 취약한 분석 결과를 보이고 있는 실정이다.

기존의 형태소 분석 시스템이 띄어쓰기에 민감한 이유는 어절 단위로 형태소 후보 리스트를 생성하고, 어절 단위로 어절 내 접속확률과 어절 간 접속확률을 이용하여 최적의 형태소 리스트를 결정하는 방법을 사용하고 있기 때문이다.

E-mail : tony@mobico.com
E-mail : albert@mobico.com

$$Pr(P) \approx \prod_i Pr(P_i \mid P_{i-1})$$

$$Pr(W \mid P) \approx \prod_i Pr(W_i \mid P_i)$$

$$P' = \arg\max_P \prod_i Pr(P_i \mid P_{i-1}) \prod_i Pr(W_i \mid P_i)$$

$$P' = \arg\max_P \sum_i \log Pr(P_i \mid P_{i-1}) + \sum_i \log Pr(W_i \mid P_i) \quad \cdots (1)$$

이 경우 사용자가 띄어쓰기를 잘못한 문자열에 대해서는 적절한 형태소 후보가 생성 조차 되지 않는 근본적 문제가 존재하게 된다. 따라서, 입력된 문자열이 모두 올바르게 띄어쓰기가 되어 있다는 전제 하에서만 올바른 형태소 분석 및 품사 태깅을 수행할 수 있었다. 이러한 문제점은 인터넷 환경 같은 불특정 다수에 의해 작성된, 띄어쓰기 오류가 많은 문서들에 대해서는 형태소 분석 및 품사 태깅 시스템의 신뢰성이 급격히 낮아지게 되고, 전술한 바와 같이 웹 문서, 게시판 등 인터넷을 통해 폭발적으로 생산되는 텍스트 정보들에 대한 분석의 한계로 작용되고 있다.

기존의 연구 및 본 연구의 목적

이와 같은 문제를 해결하기 위해서, 최근 몇 년간 형태소 분석에 있어서 띄어쓰기 보정을 위한 많은 노력이 기울여 졌다. 대부분의 연구는 형태소 분석 단계에서 다양한 휴리스틱을 사용하거나 형태소 분석의 전처리 단계에서 띄어쓰기 오류 보정을 수행하는 방법을 사용하고 있다. 휴리스틱을 사용한 띄어쓰기 보정의 경우 문장 작성자의 무한대에 가까운 오류들을 일일이 대응하기 힘들기 때문에, 일반인이 흔히 틀리기 쉬운 오류에 대해서만 일부 보정을 수행하는 한계를 가지고 있다.

주목할 만한 연구는 전처리 단계에서 띄어쓰기 오류 보정 과정을 두는 방식 중 음절 단위의 띄어쓰기 확률을 사용하는 방법인데(강승식, 1998 ; 심광섭, 1996), 이 경우 약 93% 이상의 띄어쓰기 보정률을 보이기도 하였다. 그러나, 전처리 단계에서 띄어쓰기 오류 보정률이 높지 않은 경우, 한번 보정된 결과는 다시 번복되지 않기 때문에 띄어쓰기 보정 단계에서 발생한 오류가 오히려 치명적인 분석 오류로 이어질 가능성이 높다. 특히 음절 바이그램 확률에 의한 띄어쓰기 보정의 경우 정상적으로 띄어 쓰여진 문장에 대해서도 오히려 붙여쓰기를 하거나 잘못 띄어 쓰게 되는 문제가 발생을 하고, 이는 상당히 낮은 형태소 분석 결과를 도출해 내는 한계를 보이고 있다. 자사에서 음절 띄어쓰기 확률에 의한 띄어쓰기 보정 후, 형태소 분석을 수행하는 방법을 시험한 결과 양질의 말뭉치에 대해서도 90% 이하의 형태소 분석 결과를 보였고, 전처리 과정을 거치지 않았을 경우의 분석률(98%)에 비해 심각한 성능 저하를 관찰할 수 있었다. 이 실험을 통해 확률에 의한 띄어쓰기 보정 방법이 상용화 수준이 되기 위해서는 정상적인 문서에 대해서 띄어쓰기 보정 오류가 0% 가까워야 한다는 결론을 얻을 수 있었고, 이는 모델의 특성상 근본적으로 불가능 하다고 판단된다.

따라서, 본 연구의 목적은 입력된 한국어 문자열의 띄어쓰기에 비교적 비 종속적인 확률 통계 기반 형태소 분석 및 품사 태깅 시스템의 개발에 있으며, 개발된 태깅 시스템은 띄어쓰기 오류 문장에서뿐만 아니라 정상적인 문장에 대해서는 분석 오류를 발생시키지 않아야 할 것이다.

한국어에 대해 개선된 통계적 접근

아래 Fig. 1의 a는 기존의 한국어 분석 시스템이 어절 내의 형태소를 검색하고 최적의 형태소 리스트를 결정하기 위해 접속 확률을 적용하는 방법을 보이고 있다. Fig. 1a에서 보는 것과 같이 형태소 분석을 어절 내에서만 수행함으로 띄어쓰기가 되어 있는 음절 '어'와 '학'이 동시에 포함된 형태소는 검색되지 않는다. 또한 형태소 '어'와 '학' 혹은 '어'와 '학교'는 항상 어절간 접속 확률이 적용되게 된다.

본 연구에서는 Fig. 1b와 같이 한국어 문장을 연속된 형태소의 나열로 보고, 각 형태소 간에는 어절 간 혹은 어절 내 접속 확률을 동시에 적용하되 어느 한쪽이 보다 강하게 작용될 수 있다고 가정하였다. 즉, 형태소 분석 시, 한국어

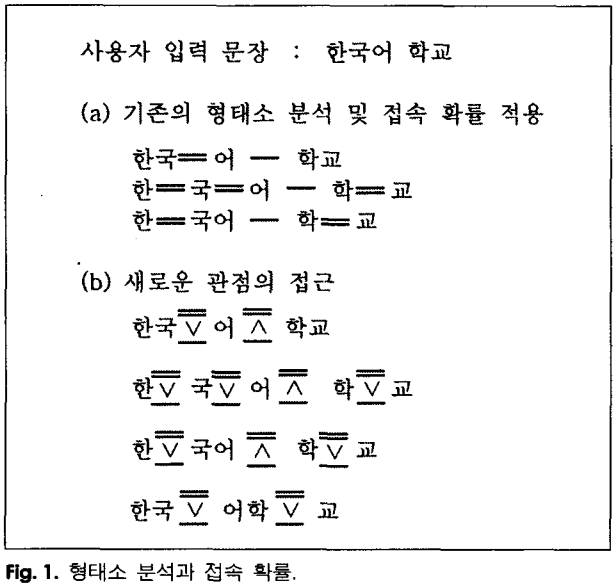

Fig. 1. 형태소 분석과 접속 확률.

를 일본어와 같이 어절 기준이 없는 언어로 가정하고 형태소 후보를 검색하며, 품사를 태깅 시에는 대규모 학습 말뭉치로부터 수집된 통계 정보와 사용자 띄어쓰기 여부, 음절 띄어쓰기 확률값을 통해 두 접속 확률의 비율이 자동 결정되도록 하자는 기본 개념에서 출발하였다.

이와 같은 개념을 HMM에 기반해 구현하고자 한다면, 말뭉치로부터의 언어 통계 자료를 구축하는 과정이 매우 중요하다. 새로운 품사 태깅 모델에서는 태깅되어 있는 대규모 학습 말뭉치로부터, 각 형태소의 출현 빈도, 품사 별 어절 내 혹은 어절 간 접속 빈도, 음절 간 띄어쓰기 확률값을 미리 확보하고 있어야 한다. 통계 기반 형태소 분석은 학습 말뭉치에서 특정 형태소 및 품사의 출현 빈도를 계산하는 것부터 시작 된다. 학습 말뭉치에의 품사 출현 빈도는 식(2)와 같이 다시 쓸 수 있고, 여기서 $freq^{CE}(P_i)$는 무시 가능하다.

$$freq(P_i) = \tilde{freq}^{\wedge}(P_i) + \tilde{freq}^{*}(P_i) + freq^{CE}(P_i)$$

$\tilde{freq}^{\wedge}(P_i)$: 어절내 형태소쌍 중 앞형태소 품사 빈도 …(2)

$\tilde{freq}^{*}(P_i)$:어절간 형태소쌍 중앞형태소 품사빈도

$freq^{CE}(P_i)$:분석 문장 전체의 마지막 품사 빈도

새로운 통계 기반 품사 태깅 모델

수식(1)은 잘 알려진 Markov의 통계 모델이고 $Pr(P_i|P_{i-1})$는 문맥확률(Contextural Probability)을, $Pr(W_i|P_i)$는 어휘확률(Lexical Probability)을 의미한다. 익히 잘 알려진 것과 같이 대규모 학습 말뭉치로부터 통계 자료를 확보한다면, 각 확률은 아래 식(3)과 같이 근사화가 가능하다.

$$\Pr(P_i \mid P_{i-1}) \approx \frac{freq(P_{i-1}, P_i)}{freq(P_{i-1})} \quad \cdots(3)$$

$$\Pr(W_i \mid P_i) \approx \frac{freq(W_i, P_i)}{freq(P_i)}$$

식(3)에 식(2)를 대입하면, 아래 식(4)와 같이 표현 가능하다, 어휘확률은 사용자 띄어쓰기 여부와 상관이 없는 값이며, 문맥확률은 식(5)와 같이 어절 내 접속 확률(Pr^)과 어절 간 접속 확률(Pr*)로 다시 나누어 표현할 수 있다.

$$\Pr(P_i \mid P_{i-1}) \approx \frac{freq(P_{i-1}, P_i)}{\tilde{freq}^{\wedge}(P_{i-1}) + \tilde{freq}^{*}(P_{i-1})} \quad \cdots(4)$$

$$\Pr(W_i \mid P_i) \approx \frac{freq(W_i, P_i)}{\tilde{freq}^{\wedge}(P_i) + \tilde{freq}^{*}(P_i)}$$

$$\Pr^{\wedge}(P_i \mid P_{i-1}) \approx \frac{freq^{\wedge}(P_{i-1}, P_i)}{\tilde{freq}^{\wedge}(P_{i-1}) + \tilde{freq}^{*}(P_{i-1})} \quad \cdots(5)$$

$$\Pr^{*}(P_i \mid P_{i-1}) \approx \frac{freq^{*}(P_{i-1}, P_i)}{\tilde{freq}^{\wedge}(P_{i-1}) + \tilde{freq}^{*}(P_{i-1})}$$

이미 앞에서 설명한 바와 같이, 기존의 품사 태깅 시스템들은 어절 내 접속 확률과 어절 간 접속 확률을 어절 경계에 따라 별도로 적용하였으나, 우리는 총 문맥확률을 어절 내 확률과 어절 간 확률의 결합으로 정의하고, 각 확률이 적정 비율로 결합된다고 가정하여, 아래 식(6)과 같이 균형 가중치 α와 β를 두었다.

$$\Pr(P_i \mid P_{i-1}) \approx \alpha \Pr^{\wedge}(P_i \mid P_{i-1}) + \beta \Pr^{*}(P_i \mid P_{i-1}) \cdots(6)$$

식(6)에 식(5)를 대입하면 아래 식(7)과 같은 어절 내 접속 확률과 어절간 접속 확률이 결합된 새로운 문맥확률 식을 얻을 수 있다.

$$\Pr(P_i \mid P_{i-1}) \approx \frac{\alpha\, freq^{\wedge}(P_{i-1}, P_i) + \beta\, freq^{*}(P_{i-1}, P_i)}{\tilde{freq}^{\wedge}(P_{i-1}) + \tilde{freq}^{*}(P_{i-1})} \cdots(7)$$

본 연구에서는 문맥 확률을 계산하기 위해 어절 내 접속 확률과 어절 간 접속 확률의 참여 비율이 학습 말뭉치로부터 추출된 음절 띄어쓰기 확률 값과 사용자의 띄어쓰기 여부의 함수로 가정하였고, 가중치 α와 β를 식(8)과 같이 정의하였다.

$$\alpha = \left(1 + \frac{\tilde{freq}^{*}(P_{i-1})}{\tilde{freq}^{\wedge}(P_{i-1}) + \tilde{freq}^{*}(P_{i-1})} \cdot \left(1 - \frac{2S \cdot V}{\gamma}\right)\right) \cdots(8)$$

$$\beta = \left(1 - \frac{\tilde{freq}^{\wedge}(P_{i-1})}{\tilde{freq}^{\wedge}(P_{i-1}) + \tilde{freq}^{*}(P_{i-1})} \cdot \left(1 - \frac{2S \cdot V}{\gamma}\right)\right)$$

S : 음절 띄어쓰기 확률(bigram) [1~].

γ : 정규화 상수 값[255].

V : 사용자 띄어쓰기 가중치[0.5 or 1.0].

형태소 분석 및 품사 태깅 시스템 개발 및 시험

제안된 품사 태깅 시스템의 개발을 위해, 본 연구에서는 자사가 확보한 세종 말뭉치 포함 총 2000만 어절에 대한 어휘 및 문맥 통계 값을 추출하였고, 분석 속도 향상을 위해 자체 개발한 trie index system을 사용하였다. 특히, 형태소 후보 과생성 방지를 위해 띄어쓰기가 확실한 어절들을 서로 묶어 pseudo 어절 단위로 분석하는 방법을 적용했으며, 다양한 휴리스틱 적용을 통해 분석 성능을 향상 시켰다.

개발된 차세대 형태소 분석기인 SMKMA는 인터넷 신문사

로부터 추출한 5000어절 규모의 시험 말뭉치에 대해 98.1%의 품사 태깅 성공률을 보였으며, 아래 시험 말뭉치와 같이 음절 단위로 모두 띄어 쓴 문장과 모두 붙여 쓴 문장 등, 극단적인 오류 상황에 있어서도 평균 90% 수준의 분석 성능을 보였다.

정상 문장 예 : 일본 국립천문대 연구팀이 태양 이외의 별 주위를 도는 행성을 발견했다고 요미우리(讀賣)신문이 4일 보도했다.....

공백 삭제문 예 : 일본국립천문대연구팀이태양이외의별주위를도는행성을발견했다고요미우리(讀賣)신문이4일보도했다.....

공백 삽입문 예 : 일 본 국 립 천 문 대 연 구 팀 이 태 양 이 외 의 별 주 위 를 도 는 행 성 을 발 견 했 다 고 요 미 우 리 (讀 賣) 신 문 이 4 일 보 도 했 다 .....

아래 Fig. 2는 2003년 9월 3일자 연합신문 기사 5편(총 591어절)에 대해 자사의 SMKMA와 국내에서 잘 알려진, K 형태소 분석기와 H 형태소 분석기 최신 버전을 사용한 시험 결과 비교를 보이고 있다.

K 형태소 분석기와 H 형태소 분석기도 띄어쓰기 오류 보정 기능을 가진 것으로 설명되고 있으나, 위와 같이 극단적

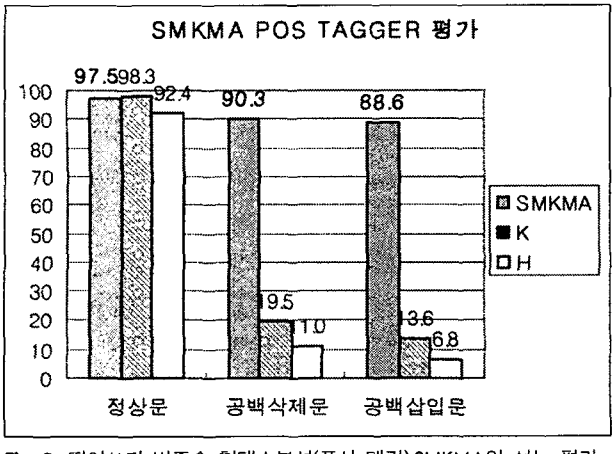

Fig. 2. 띄어쓰기 비종속 형태소분석(품사 태깅) SMKMA의 성능 평가.

인 띄어쓰기 오류에 대해서는 매우 낮은 분석 성공률을 보이고 있고, 새로운 모델이 적용된 자사의 SMKMA의 경우 모든 경우에 있어서 상당히 높은 분석 성능을 가졌음을 확인할 수 있었다.

결 론

본 논문에서는 개선된 HMM 기반의 형태소 분석 및 품사 태깅 모델을 제시하였고, 실제 개발된 시스템의 성능 시험을 통해, 시스템이 사용자 띄어쓰기에 비교적 무관하게 분석을 수행해 내고 있음을 확인할 수 있었다. 새로운 품사 태깅 시스템은 분석 속도 개선을 위해 자체 제작된 고성능 trie index system과 pseudo 어절 처리 알고리즘을 채용했으며, 분석 성능 향상을 위해서 다양한 휴리스틱의 적용과 함께 기분석 사전을 활용 하였다. 개발된 SMKMA는 총 40만 한국어 형태소 사전, 5만개 이상의 격틀 정보, 3000종으로 분류된 30만 단어 이상의 시소러스 사전을 포함하고 있다.

SMKMA는 현재 모비코 & 시스메타(MnS)의 문서 개체 검색 시스템인 [IN2] DOR, 텍스트 마이닝 시스템인[IN2] TMS, 그리고 차세대 다국어 자동번역 시스템인 Transwiz (번역마법사)의 하부 엔진으로 사용되고 있고, 관련 기술이 특허로 출원되어 있다.

REFERENCES

김진동, 이상주, 임해창(1998) : "어절 띄어쓰기를 고려한 형태소 단위 품사 태깅 모델", 제 10 회 한글 및 한국어 정보처리 학술대회

강승식(1998) : "한글 문장의 자동 띄어쓰기", 제 10 회 한글 및 한국어 정보처리 학술발표 논문집, *pp137-142*

김홍규 외(2000) : "21세기 세종계획 국어 기초자료 구축 연구 보고서", 문화관광부

심광섭(1996) : "음절간 상호 정보를 이용한 한국어 자동 띄어쓰기", 정보과학회논문지 제 23 권

신상현, 이근배, 이종혁(1997) : "통계와 규칙에 기반한 2단계 한국어 품사 태깅 시스템", 정보과학회논문지 제 24 권

김재훈, 임철수, 서정연(1995) : "은닉 마르코프 모델을 이용한 효율적인 한국어 품사 태깅", 한국정보과학회논문지 22권