띄어쓰기 오류를 고려한 최적 형태소 분석을 위한 확률 모델

최적 형태소 분석 문제는 주어진 문장 S에 대해 가능한 모든 형태소 분석 후보들 중에서 주어진 문장에 대한 출현 확률 $P(M|S)$이 가장 높은 분석 후보 $\widehat{M}$을 찾는 것이라 할 수 있고, 이는 식(1)과 같이 표현될 수 있다.

$$\widehat{M} = \underset{M}{argmax}\, P(M|S) \tag{1}$$

만약 해당 문장이 띄어쓰기 오류를 포함하고 있다면, 이는 띄어쓰기 오류 교정과 형태소 분석을 동시에 해결하는 문제가 된다. 이때 띄어쓰기 오류 교정 문제는 <그림1>에서와 같이 주어진 문장에 대해서 각 음절 사이의 띄어쓰기 오류 여부를 태깅 하는 것으로 생각할 수 있다.

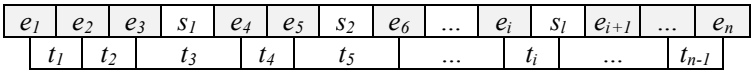

그림 1. 띄어쓰기 오류 태깅

따라서, 형태소 분석 문제를 띄어쓰기 오류를 포함한 문장으로 확장하면, 주어진 문장에 대해서 띄어쓰기 오류 교정 태깅과 형태소 분석 확률을 최대화 하는 형태소 분석 후보를 찾는 것으로 변경할 수 있고, 식(1)은 식(2)와 같이 변경된다.

$$\widehat{M} = \underset{M}{argmax}\, P(M, T_s|S) \tag{2}$$

이때 띄어쓰기 교정과 형태소 분석의 출현 확률 $P(M, T_s|S)$는 식(3)에서와 같이 주어진 문장에 대한 띄어쓰기 교정 태깅 확률 $P(T|S)$와 띄어쓰기가 교정된 문장에 대한 형태소 분석 출현 확률 $P(M|T, S)$의 곱으로 표현할 수 있다.

$$P(M, T_s|S) = P(T_s|S)P(M|T_s, S) \tag{3}$$

여기서 S는 띄어쓰기를 포함한 음절의 연속으로 $S = [e_1 e_2 e_3 s_1 e_4 e_5 s_2 \dots e_i s_j \dots e_n]$와 같이 표현된다. 여기서 e_i는 공백이 아닌 i번째 음절을 의미하고 s_j는 j번째 공백을 의미한다. 띄어쓰기에 대한 태깅 T_s는 $T_s = [t_1 t_2 \dots t_i \dots t_{n-1}]$와 같이 표현된다. 이때 $t_i \in \{true, false\}$이다. 음절간의 띄어쓰기가 옳게 되었다고 판단된 경우에는 *true*를 그르게 되었다고 판단된 경우에는 *false*를 지정하게 된다. 따라서 공백을 낀 경우와 공백을 끼지 않은 경우로 나뉘어서 <그림1>과 같이 음절의 수보다 하나 적은 수만큼 교정 태그가 붙여진다. 이를 이용하여 $P(T_s|S)$를 확장하면 식(4)와 같이 표현된다.

$$\begin{aligned}
P(T_s|S) &= P(t_1 t_2 \dots t_i \dots t_{n-1}|S) \\
&= P(t_1|S)\, P(t_2 \dots t_i \dots t_{n-1}|t_1, S) \\
&= P(t_1|S)\, P(t_2|t_1, S) P(t_3 \dots t_i \dots t_{n-1}|t_1 t_2, S) \\
&= P(t_1|S)\, P(t_2^s|t_1^s, S) \dots P(t_i|t_1 t_2 \dots t_{i-1}, S) \dots P(t_{n-1}|t_1 t_2 \dots t_i \dots t_{n-2}, S) \\
&= P(t_1|S) \prod_{i=2}^{n-1} P(t_i|t_1 \dots t_{i-1}, S)
\end{aligned} \tag{4}$$

이때, 음절 e_i에서의 띄어쓰기 태깅 확률 $P(t_i|t_1 t_2 \dots t_{i-1}, S)$값을 구하는 것은 현실적으로 불가능하기 때문에, i번째 띄어쓰기 교정 태깅 t_i는 직전 음절 e_i와 다음 음절 e_{i+1}에 의해서만 영향을 받는다고 가정하면, 식(4)는 식(5)와 같이 단순화 할 수 있다.

$$P(T_s|S) = \prod_{i=1}^{n-1} P(t_i|e_i e_{i+1}) \tag{5}$$

띄어쓰기 교정 태그 T_s가 문장 S에 적용된 것을 $\bar{S}$이라 하면 $\bar{S}$은 α개의 어절의 연속으로 $\bar{S} = [E_1 \dots E_k \dots E_\alpha]$와 같이 표현되고 $P(M|T_s, S)$는 $P(M|\bar{S})$로 표현된다. 형태소 분석을 각 어절 단위로 구분하여 표현하면, 형태소 분석 결과 M은 어절 E_k에서의 형태소 분석 결과 M_k의 연속으로 $M = [M_1 \dots M_k \dots M_\alpha]$와 같이 표현된다. 이를 이용하여 $P(M|\bar{S})$를 확장하면 식(6)과 같이 표현된다.

$$\begin{aligned} P(M|\bar{S}) &= P(M_1 \dots M_\alpha | E_1 \dots E_\alpha) \\ &= P(M_1|E_1 \dots E_\alpha) \prod_{k=2}^{\alpha} P(M_k|M_1 \dots M_{k-1}, E_1 \dots E_\alpha) \end{aligned} \tag{6}$$

$P(M_k|M_1 \dots M_{k-1}, E_1 \dots E_\alpha)$를 구하는 것은 현실적으로 불가능 하기 때문에 M_k의 의존성을 단순화하여 고려할 필요가 있다. 일반적으로 한국어에서의 어순은 자유로우나, 관형어 다음에는 명사가 오는 것과 같이, 특별한 경우는 어순에 제약이 따른다. 따라서 이를 반영하여 M_k는 E_k와 M_{k-1}에 의해서 영향을 받는다고 가정한다. 이를 바탕으로 식(6)은 식(7)과 같이 단순화된다.

$$P(M|\bar{S}) = \prod_{k=1}^{\alpha} P(M_k|M_{k-1}, E_k) \tag{7}$$

여기서 M_0는 문장의 시작을 나타내는 특별한 형태소 분석 결과로 $M_0 = [(null, \text{^})]$이다. M_k는 λ_k개의 형태소의 연속이라고 하면 $M_k = [(m_{k,1}, p_{k,1}) \dots (m_{k,\mathrm{u}}, p_{k,\mathrm{u}}) \dots (m_{k,\lambda}, p_{k,\lambda_k})]$와 같이 표현된다. 여기서 $m_{k,u}$는 형태소를 나타내고, $p_{k,\mathrm{u}}$는 형태소 $m_{k,\mathrm{u}}$의 범주를 나타낸다. Γ는 모든 가능한 형태소의 집합이고, Δ는 가능한 형태소 범주의 집합으로 정의한다.

$P(M_k|M_{k-1}, E_k)$를 이를 이용하여 확장하면 식(8)과 같이 표현된다.

$$\begin{aligned} P(M_k|M_{k-1}, E_k) &= P\left((m_{k,1}, p_{k,1}) \dots (m_{k,u}, p_{k,u}) \dots (m_{k,\lambda_k}, p_{k,\lambda_k}) | M_{k-1}, E_k\right) \\ &= P\left((m_{k,1}, p_{k,1}) | M_{k-1}, E_k\right) \prod_{u=2}^{\lambda_k} P\left((m_{k,u}, p_{k,u}) | (m_{k,1}, p_{k,1}) \dots (m_{k,u-1}, p_{k,u-1}), M_{k-1}, E_k\right) \end{aligned} \tag{8}$$

각각의 형태소 분석 후보 M_k가 가변 길이의 매우 다양한 조합을 만들어내는 것을 감안한다면, 여전히 현실적으로 이 같은 확률 값을 얻어내는 것은 거의 불가능하다. 또한, 형태소의 수를 감안한다면, 형태소의 bigram을 이용하는 것도 거의 불가능 하다. 따라서 어절 내의 형태소 분석 결과는 이전 형태소의 범주에만 영향을 받는다고 가정한다. 이 같은 가정으로부터 식(8)은 식(9)와 같이 단순화 된다. 여기서 $p_{k,0}$는 이전 형태소 분석 결과의 마지막 형태소, 즉 $p_{k-1,\lambda_{k-1}}$이다.

$$P(M_k|M_{k-1}, E_k) = \prod_{u=1}^{\lambda_k} P(m_{k,u}, p_{k,u}|p_{k,u-1}) \tag{9}$$

$P(m_{k,u}, p_{k,u}|p_{k,u-1})$는 베이스 정리에 의해서 식(10)과 같이 변형된다.

$$P(m_{k,u}, p_{k,u}|p_{k,u-1}) = \frac{P(p_{k,u-1}|m_{k,u}, p_{k,u})P(m_{k,u}, p_{k,u})}{P(p_{k,u-1})} \tag{10}$$

식(10)을 식(9)에 적용하고, 식(9)를 식 (7)에 적용하면 식(7)은 식(11)과 같이 단순화된다.

$$P(M|T_s, S) = P(M|\bar{S}) = \prod_{k=1}^{\alpha}\prod_{u=1}^{\lambda_k} \frac{P(p_{k,u-1}|m_{k,u}, p_{k,u})P(m_{k,u}, p_{k,u})}{P(p_{k,u-1})} \tag{11}$$

식(5)와 식(11)을 식(3)에 적용하면 $P(M, T_s|S)$은 식(12)와 같이 다시 표현된다.

$$P(M, T_s|S) = \prod_{i=1}^{n-1} P(t_i|e_i e_{i+1}) \times \prod_{k=1}^{\alpha}\prod_{u=1}^{\lambda_k} \frac{P(p_{k,u-1}|m_{k,u}, p_{k,u})P(m_{k,u}, p_{k,u})}{P(p_{k,u-1})} \tag{12}$$

따라서 형태소 분석을 수행할 때 필요한 사전 확률은 $P(t_i|e_i e_{i+1})$, $P(p_{k,u-1}|m_{k,u}, p_{k,u})$, $P(m_{k,u}, p_{k,u})$ 및 $P(p_{k,u-1})$가 된다. 여기서 $P(p_{k,u-1})$는 $|\Delta|$가 크지 않으므로 학습에 문제가 되지 않는다. 그러나 나머지 세 값들은 학습 데이터가 충분하지 않을 경우 그 값이 발견되지 않거나 왜곡될 수 있기에 이를 보완할 수 있는 평탄화 작업(smoothing)이 필요하다.

$P(p_{k,u-1}|m_{k,u}, p_{k,u})$의 경우 $count(m_{k,u}, p_{k,u})$가 $|\Delta|$보다 매우 큰 경우 신뢰할 수 있는 값을 얻을 수 있다. 따라서 $P(p_{k,u-1}|m_{k,u}, p_{k,u})$는 식(13)과 같이 평탄화 할 수 있다.

$$P(p_{k,u-1}|m_{k,u}, p_{k,u}) = \begin{cases} P(p_{k,u-1}|m_{k,u}, p_{k,u}), & if\ count(m_{k,u}, p_{k,u}) \gg |\Delta| \\ P(p_{k,u-1}|p_{k,u}), & otherwise \end{cases} \tag{13}$$

학습데이터에서 발견되지 않은 $m_{k,u}, p_{k,u}$에 대해서는 $P(m_{k,u}, p_{k,u})$을 $\min_{m\in\Gamma, p\in\Delta} P(m, p)$로 하고, 발견되지 않는 음절 조합에 대해서 $P(t_i|e_i e_{i+1})$를 0.5로 한다. 이 같은 평탄화를 이용하면, 주어진 학습데이터 $\mathcal{D}$를 바탕으로 $P(M, T_s|S)$를 추정할 수 있다. 여기서 학습데이터 $\mathcal{D}$는 바르게 띄어쓰기 되었고, 각 어절에 대한 형태소 분석이 된 문장의 집합이다.