Active Learning과 군집화를 이용한 고정키어구 추출

이현우(창원대), 차정원(창원대)

<차 례>

<Abstract>

Keyphrase Extraction Using Active Learning and Clustering

Hyun-Woo Lee, Jeong-Won Cha

We describe a new active learning method in conditional random fields (CRFs) framework for keyphrase extraction. To save elaboration in annotation, we use diversity and representative measure. We select high diversity training candidates by sentence confidence value. We also select high representative candidates by clustering the part-of-speech patterns of contexts. In the experiments using dialog corpus, our method achieves 86.80% and saves 88% training corpus compared with those of supervised method. From the results of experiment, we can see that the proposed method shows improved performance over the previous methods. Additionally, the proposed method can be applied to other applications easily since its implementation is independent on applications.

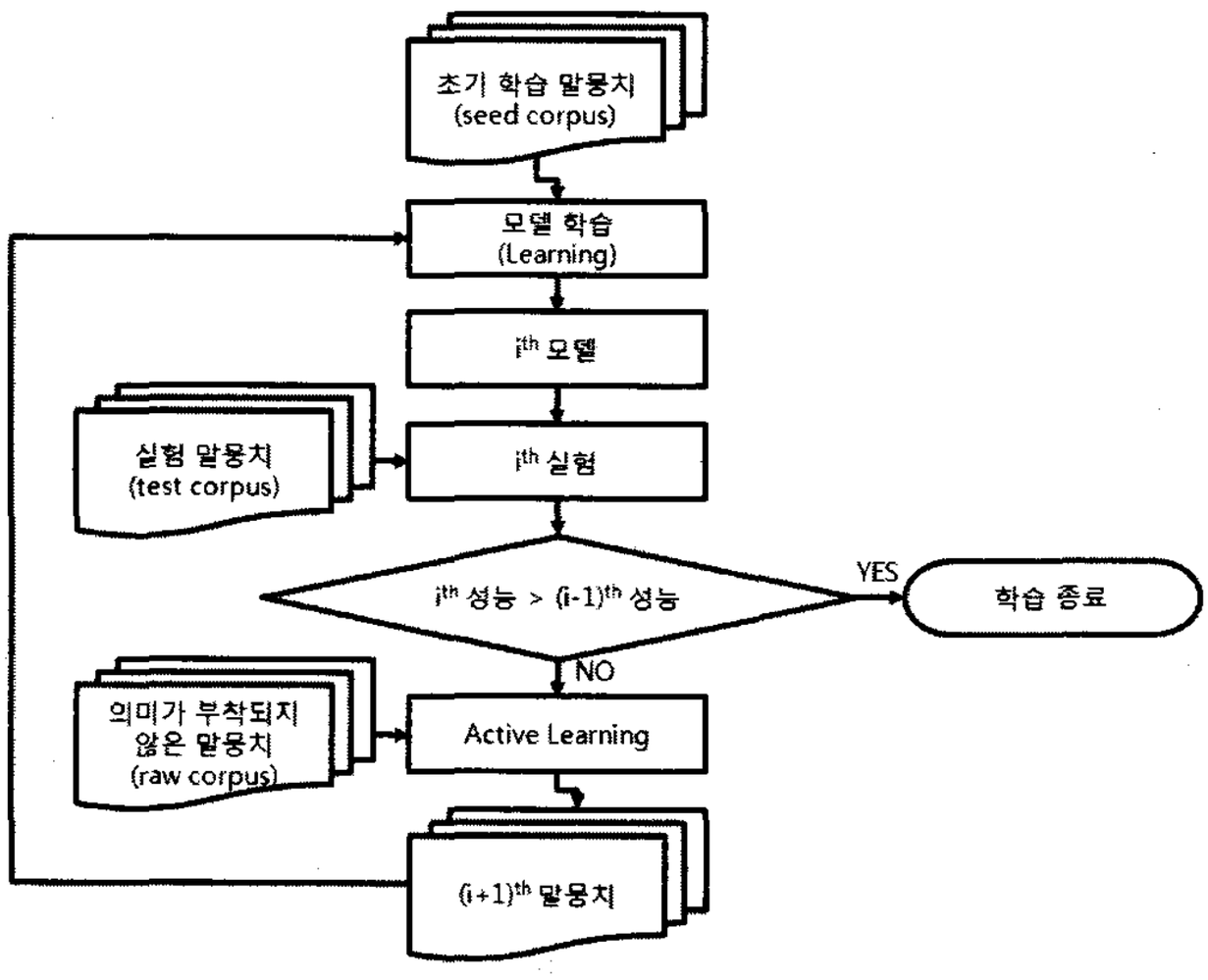
* Keywords: Machine learning, Conditional random fields (CRFs), Active learning, Keyphrase, Clustering, Representativeness, Diversity.

1. 서 론

고정키어구란 '하나 이상의 키워드가 어구를 이루어 일정한 영역에서 고정된 의미로 사용되는 어구(phrase)'로 정의한다. 예를 들면, "하늘이시여", "착한 여자 나쁜 여자", 그리고 "TV는 사랑을 싣고" 등이다. 고정키어구를 추출하기 위해서는 미리 언어정보 부착 작업이 완료된 학습용 말뭉치를 기계학습을 하는 방법이 주류를 이룬다.

기계학습의 가장 큰 문제점은 언어정보가 부착된 많은 양의 학습용 말뭉치가 필요하다는 점이다. 학습용 말뭉치의 언어정보 부착 작업[1]은 막대한 인적, 물적 자원을 필요로 한다. 또한, 작업 자체가 어렵고 까다로우며, 많은 사람이 함께 하는 작업이므로 '잘못된 언어정보'가 부착되는 문제점을 가지고 있다.

본 논문에서는 위와 같은 어려운 언어정보 부착 작업을 적은 양의 학습용 말뭉치와 active learning[1]을 이용하여 이러한 문제를 보다 효율적으로 해결하고자 한다. Active learning은 <그림 1>과 같이 "초기 학습용 말뭉치"(seed corpus)를 이용하여 모델을 학습하고, 이 모델을 이용하여 "언어정보가 부착되지 않은 말뭉치"(raw corpus)에 언어정보를 부착한 뒤, 일부를 사용자에게 다시 언어정보 부착 작업을 요청한다. 여기서 사용자는 모델이 부착한 언어정보를 수정하게 되며, 이를 학습용 말뭉치에 추가하여 다음 학습에 이용하는데, 이러한 과정을 사용자가 만족하는 성능에 도달할 때까지 반복하게 된다.



<그림 1> Active learning 방법

1) 여기서 언어정보 부착 작업은 품사 부착, 구문 정보 부착 등 자연어처리에서 사용되는 모든 정보를 부착하는 작업을 포함하는 말로 사용한다.

2. 관련 연구

Active learning에서는 사용자에게 언어정보 부착 작업을 요청할 때, 성능을 많이 증가시킬 수 있는 말뭉치를 전달하는 것에 초점을 맞추어 연구가 이루어 졌다 [2]-[5].

문서 분류기를 학습할 때, 주어진 문서를 양성예제(positive example)와 음성예제(negative example)로 구별하는데, Naive Bayes 알고리즘으로 양성예제에 속할 확률이 0.5인 예제, 즉 가장 구별하기 힘든 예제만을 선정하여 전문가에게 올바른 문서 분류를 선택한 다음, 학습에 사용된다 [2]. 앞의 연구와 비슷하게 학습 단계에서 다수의 위원회(committee)를 두어, 위원들 사이에서 동일한 예제에 대하여 각 위원간의 의견이 가장 일치하지 않은 예제만을 선정하여 올바른 언어정보를 부착한 다음, 학습에 이용한다 [3]. 이는 [2]와 [3]에서 불확실성이 높은 예제들이 성능을 가장 빨리 올릴 수 있는 의미있는 예제라고 정의하였기 때문이다.

개체명 부착기를 학습할 때, [2]나 [3]과 같이 불확실성이 높은 의미 있는 예제뿐만 아니라, 동일하거나 비슷한 예제가 학습되는 것을 방지하기 위해서 개체명 또는 단어 사이의 유사도와 kNN-clustering으로 대표적(representativeness)이면서 다양한 예제(diversity)를 선정하여 학습하였으며, 그 결과 supervised learning보다 학습양이 약 80%정도 감소하였다 [4].

화자의 의도를 파악하기 위한 의미구조 분석기를 학습할 때, 학습에 사용된 예제 중에 conditional random fields (CRFs)의 정확도가 가장 낮은 예제를 선정하여 올바른 예제로 수정하고 학습을 하였는데, 이는 CRFs의 정확도가 낮다는 것은 CRFs로 생성된 모델이 표현할 수 없는 예제라고 정의하였기 때문이다. 그 결과, 약 56%정도 학습량이 감소하였다 [5].

3. CRFs를 이용한 고정키어구 추출

본 논문에서는 고정키어구 추출을 품사 부착 문제와 같이 형태소에 바로 태그(tag)를 붙이는 방식으로 해결하였다. "형태소/품사/$\{B,I,O\}$"형식으로 구성된 예를 보면 "나/NP/B 는/JX/O 착하/VA/B ㄴ/ETM/I 여자/NNG/I 나쁘/VA/I ㄴ/ETM/I 여자/NNG/I 를/JKO/O 보/VV/O 았/EP/O 다/EF/O ./SF/O"와 같이 고정키어구가 $\{B,I,O\}$로 표기된다.

<그림 2>에서는 먼저 논문에 사용된 용어에 대해 정의한다. X는 문장을 구성하는 형태소 x들의 벡터, Y는 문장을 구성하는 품사 y들의 벡터, 문장에서 고정키어구를 나타내는 T는 $\{B,I,O\}$로 구성된 벡터로 정의된다.

$X=< x_1, x_2, x_3, \cdots, x_n >$
$Y=< y_1, y_2, y_3, \cdots, y_n >$
$T=< t_1, t_2, t_3, \cdots, t_n >$
$t=\{B,I,O\}$
B: 고정키어구의 시작
I: 고정키어구의 중간 또는 끝
O: 고정키어구 아님

<그림 2> 용어 정의

3.1. Conditional Random Fields

CRFs는 조건부 확률을 최대로 하는 방향성이 없는 그래프 모델이다 [6]. 입력열 $X=x_1x_2\cdots x_n$, 상태열 $T=t_1t_2\cdots t_n$가 주어지고 가중치 $\Lambda=\{\lambda_1,...\}$가 주어졌을 때, CRFs에서는 조건 확률로 식 (1)과 같이 정의된다.

$$P(T|X)=\frac{1}{Z(X)}exp\left(\sum_{i=1}^{n}\sum_{k}\lambda_k f_k(t_{i-1},t_i,x,i)\right) \quad (1)$$

여기서 $Z(X)$는 확률값으로 만들어 주는 정규화 값이고 $f_k(t_{i-1},t_i,x,i)$는 자질 함수이다. 또한 λ_k는 각 자질에 대한 가중치를 나타낸다. k는 k번째 자질이며, 자질 함수는 현재 시간 i에 대해 관측열 x, 상태변이 $t_{i-1}\rightarrow t_i$에 대해서 전이의 양상을 측정할 수 있다.

매개변수들은 주어진 입력열과 이에 대응하는 상태열에 대한 조건부 확률이 최대화하는 최대 유사도(maximum likelihood)에 의해서 추정된다. 훈련 집합 $\{(t_l,x_l)\}_{l=1}^{N}$에 대해서 다음과 같은 로그 유사도(log-likelihood)를 계산한다.

$$L(\Lambda)=\sum_{l}\log P_\Lambda(t_l|x_l)=\sum_{l}\left(\sum_{i=1}^{n}\sum_{k}\lambda_k f_k(t_{i-1},t_i,x,i)-\log Z_{x_i}\right) \quad (2)$$

CRFs는 식 (2)를 최대로 하도록 학습한다. 일반적으로 CRFs는 improved iterative scaling (IIS)나 generalized iterative scaling (GIS)[7]를 사용하여 학습한다. 또한 학습 데이터의 과적합(overfitting) 문제를 해결하기 위해서 가우스 사전 평활(Gaussian prior smoothing)[8]을 적용한다.

3.2. 고정키어구 추출 모델의 자질 정의

고정키어구 추출 모델의 자질을 정의한 <표 1>은 현재 형태소/품사로부터 최대

앞으로 두 형태소/품사, 뒤로 두 형태소/품사를 사용하게 되어 있다. 이유는 [9]에서 uni-gram과 four-gram보다는 bi-gram과 tri-gram이 더 좋은 성능을 보인다고 하였기 때문이다.

<표 1> 고정키어구 추출 모델의 자질 정의, x_i는 주어진 문장의 현재 형태소, y_i는 주어진 문장의 현재 품사를 가리킨다. 예제는 "나/NP 는/JX 착하/VA ㄴ/ETM 여자/NNG 나쁘/VA ㄴ/ETM 여자/NNG 를/JKO 보/VV 았/EP 다/EF ./SF"에서 현재 형태소가 "착하"일 때의 예제.

자질 번호	자질정의	예제	자질 번호	자질정의	예제
0	x_{i-2}	나	13	$x_{i-1}/y_{i-1}/x_i/y_i$	는/JX/ 착하/VA
1	x_{i-1}	는	14	$x_i/y_i/x_{i+1}/y_{i+1}$	착하/VA/ 는/JX
2	x_i	착하	15	y_{i-2}	NP
3	x_{i+1}	ㄴ	16	y_{i-1}	JX
4	x_{i+2}	여자	17	y_i	VA
5	$x_{i-2}/x_{i-1}/x_i$	나/는/착하	18	y_{i+1}	ETM
6	$x_{i-1}/x_i/x_{i+1}$	는/착하/ㄴ	19	y_{i+2}	NNG
7	$x_i/x_{i+1}/x_{i+2}$	착하/ㄴ/여자	20	$y_{i-2}/y_{i-1}/y_i$	NP/JX/VA
8	x_{i-1}/x_i	는/착하	21	$y_{i-1}/y_i/y_{i+1}$	JX/VA/ETM
9	x_i/x_{i+1}	착하/ㄴ	22	$y_i/y_{i+1}/y_{i+2}$	VA/ETM/NNG
10	x_{i-1}/y_{i-1}	는/JX	23	y_{i-1}/y_i	JX/VA
11	x_i/y_i	착하/VA	24	y_i/y_{i+1}	VA/ETM
12	x_{i+1}/y_{i+1}	ㄴ/ETM			

3.3. Active Learning을 이용한 고정키어구 모델 학습

본 장에서는 고정키어구를 추출하기 위해 active learning으로 학습하는데, 사용된 말뭉치 정보 및 평가 방법, 그리고 대표 학습 후보 선정에 대해 기술한다.

3.3.1. 학습 및 실험용 말뭉치

본 논문에서는 대화 시스템을 위해 제작된 총 5,310 문장(48,503 형태소)의 말뭉치[10]를 사용하였으며, 품사부착기의 오류가 실험에 미치는 영향을 최소화하기 위해서 모든 품사 오류를 수정하였다. 지도학습에서는 5,010 문장을 학습에 사용하고 300 문장을 평가에 사용하였다. 본 논문에서 제안할 active learning을 위해서 동일한 문서를 초기 학습용 말뭉치 10 문장, 언어정보가 부착되지 않은 말뭉치

5,000 문장, 평가용 말뭉치 300 문장으로 나누었다. 초기 학습용 말뭉치는 초기 고정키어구 추출 모델을 생성하기 위해 사용하였다.

<표 2> 학습 및 실험 말뭉치 정보

	초기 학습용 말뭉치	언어정보가 부착되지 않은 말뭉치	평가용 말뭉치
문장 수	10	5,000	300
형태소 수	156	46,035	2,311
평균 형태소 수	15	9	7
평균 고정키어구 수	2	2	2
평균 고정키어구 길이	5	2	2

3.3.2 평가 기준

본 논문에서는 고정키어구 평가 기준을 고정키어구 정확도(P_{key}), 고정키어구 재현률(R_{key})을 사용한다. 단, 고정키어구의 시작과 끝이 정확하게 일치할 때만 고정키어구 정답으로 인정한다. 그리고 성격이 다른 정확도와 재현률을 조합하여 전체적인 성능을 나타내기 위해 F-measure(F_{key})도 평가 기준으로 추가하였다.

$$P_{key} = \frac{\text{올바르게 추출한 고정키어구의 개수}}{\text{고정키어구 모델이 추출한 고정키어구의 개수}} \times 100(\%), \quad (3)$$

$$R_{key} = \frac{\text{올바르게 추출한 고정키어구의 개수}}{\text{평가용 말뭉치의 고정키어구 수}} \times 100(\%), \quad (4)$$

$$F_{key} = \frac{(\beta+1) \times P_{key} \times R_{key}}{\beta \times P_{key} + R_{key}}, \beta = 1. \quad (5)$$

3.3.3. 학습할 문장 선정 방법

[5]에서는 문장 전체의 CRFs 정확도가 낮은 문장만을 선정하였는데, 본 논문에서는 고정키어구만의 문장신뢰도[1]를 계산하는 식 (6)을 사용하여, 고정키어구가 많이 출현하거나, 긴 고정키어구가 출현한 문장 중에 "문장신뢰도[1]이 높은 상위 50개", "문장신뢰도[1]이 낮은 하위 50개"를 선정하였다.

여기서 KP_j는 문장에서 j번째 고정키어구를 나타내고 $x_{j,1}$은 j번째 고정키어구에서 첫 번째 형태소를 나타내며, h_j는 j번째 고정키어구의 길이, c는 문장에 나타난 고정키어구 갯수를 나타낸다. $P(x_{j,r})$은 CRFs가 출력하는 형태소 $x_{j,r}$의 확률값이다.

$$KP_j = \langle x_{j,1}, x_{j,2}, x_{j,3}, \cdots, x_{j,h_j} \rangle,$$

$$h_j = |KP_j|,$$

$$c = \text{한 문장에서 고정키어구 갯수}, \tag{6}$$

$$\text{문장신뢰도}^1 = \frac{1}{c}\sum_{j=1}^{c}\left[\frac{1}{h_j^2}\sum_{r=1}^{h_j} P(x_{j,r})\right].$$

Active learning을 중단하는 시점은 <표 2>의 언어정보가 부착되지 않은 말뭉치에 언어정보를 부착하여 초기 학습용 말뭉치와 합친 5,010문장의 말뭉치를 이용해서 지도학습(supervised learning)으로 학습한 모델의 성능 <표 3>을 능가하거나, 성능(F_{key}) 증가가 3회 동안 없을 경우이다.

<표 4>와 <그림 3>의 "실험 1"은 학습 말뭉치에 과적합된, 즉 초기 학습 말뭉치와 비슷한 문장들만 선정되는 문제가 있어, 성능이 74.28%에서 학습을 멈추었다. 반면, "실험 2"는 지도학습의 성능인 <표 3>의 성능에는 못 미치지만 "실험 1"보다는 월등히 높은 성능인 83.31%를 기록하였다. 이는 초기 학습용 말뭉치와는 전혀 달라서 성능을 증가시킬 수 있는 다양한 문장들이 선정되었기 때문이다. 즉 하위 50개의 문장이 다양성(diversity)을 잘 표현하고 있다고 할 수 있다. 그럼에도 불구하고 <표 3>보다 성능이 낮은 이유는 성능을 증가시킬 수 있는 문장들을 학습하지 못했기 때문이다.

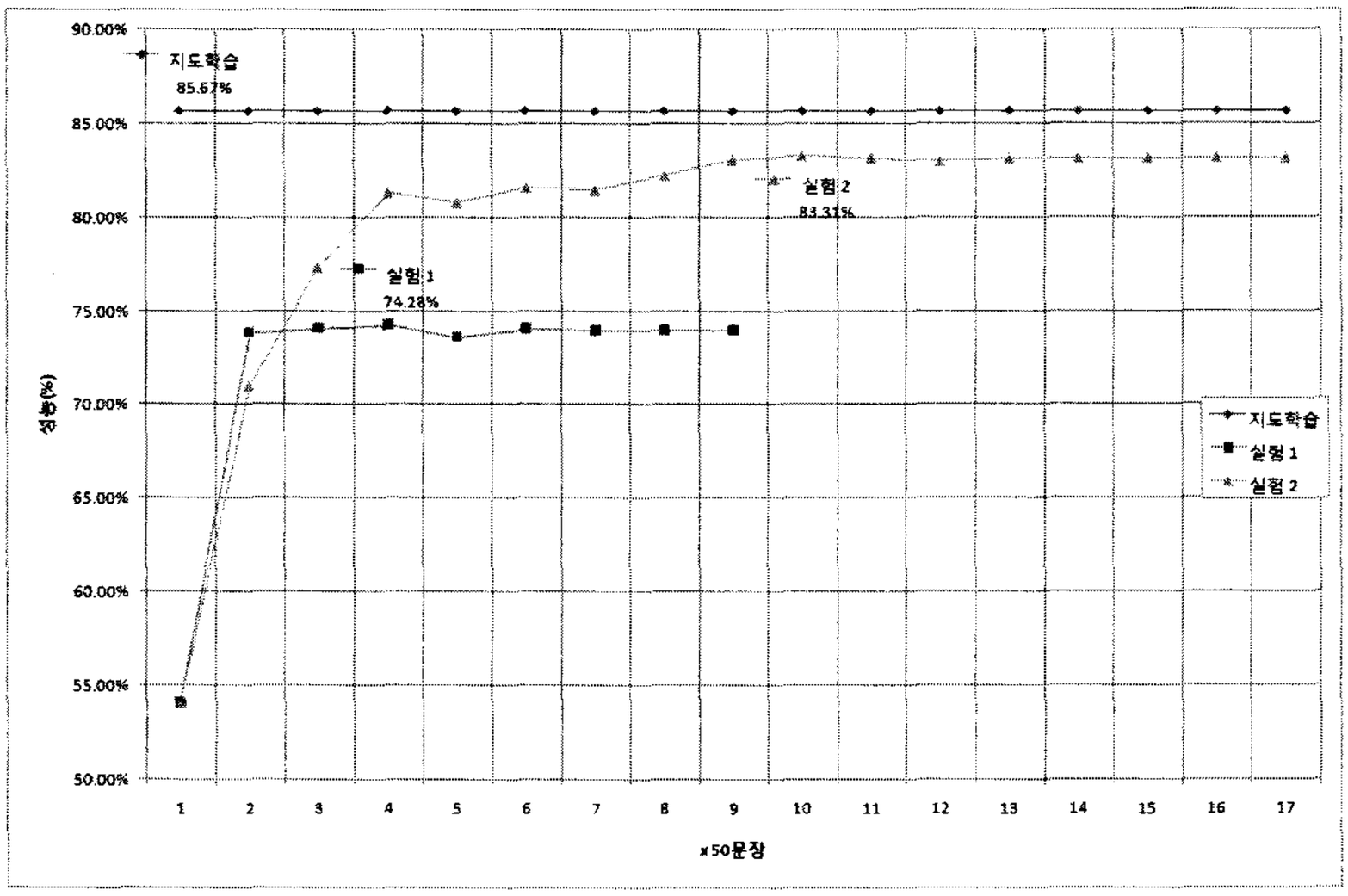
<표 3> 지도학습한 고정키어구 모델의 성능, 단위(%)

P_{key}	R_{key}	F_{key}
89.28	82.35	85.67

<표 4> "실험 1, 2"의 문장 선정 조건 및 성능(F_{key}) 표

실험 번호	문장 선정 조건	최적 성능[2]		최고 성능	
		단계	성능	단계	성능
실험 1	문장신뢰도[1] 상위 50개	-	-	4	74.28%
실험 2	문장신뢰도[1] 하위 50개	-	-	10	83.31%

2) 최적 성능은 지도학습의 성능을 넘어선 시점의 성능을 말한다.



<그림 3> 지도학습 및 "실험 1, 2"의 성능 그래프, F_{key}기준

3.3.4. 대표 학습 후보 선정

본 논문에서는 다양하면서 대표성(representativeness)을 가진 문장를 선정하기 위해, 품사패턴을 이용한 군집화를 사용한다. 이것은 CRFs 학습에서 비슷한 문맥을 가질 경우, 비슷한 결과를 나타내는 경향이 많기 때문에 이를 군집화하여 대표 품사패턴만을 사용한다. 품사패턴은 고정키어구와 고정키어구 주위에 있는 품사들의 패턴이며, 수집된 품사패턴 중에 문장신뢰도[1]이 낮은 하나의 문장만을 선정해서 매 학습마다 일정 개수의 패턴을 학습하도록 한다. 식 (7)은 품사패턴의 유사도를 계산하는 식이며, P_u와 P_v는 임의의 두 키어구 품사 문맥 단락이다.

$$size = 2$$

$$P_u = (y_{u,1}, y_{u,2}, y_{u,3}, \cdots y_{u,n}),$$
$$P_v = (y_{v,1}, y_{v,2}, y_{v,3}, \cdots y_{v,m}),$$
$$n < m,$$

$$s(y_{u,q}, y_{v,q}) = \begin{cases} 1 & \text{if}(y_{u,q} = y_{v,q}), \\ 0 & otherwise, \end{cases} \tag{7}$$

품사패턴유사도(P_u, P_v)

$$= \frac{1}{2 \times \left\lfloor \frac{n}{2} \right\rfloor} \sum_{q=1}^{\left\lfloor \frac{n}{2} \right\rfloor} \left(s(y_{u,q}, y_{v,q}) + s(y_{u,n-q-1}, y_{v,m-q-1}) \right).$$

여기서 size는 고정키어구 앞뒤 문맥의 크기이며, m과 n은 고정키어구의 품사패턴 길이를 각각 나타낸다.

두 문장 "나/NP/B 는/JX/O 착하/VA/B ㄴ/ETM/I 여자/NNG/I 를/JKO/O 보/VV/O 았/EP/O 다/EF/O ./SF/O"와 "나/NP/B 는/JX/O 착하/VA/B ㄴ/ETM/I 여자/NNG/I 나쁘/VA/I ㄴ/ETM/I 여자/NNG/I 를/JKO/O 녹화/NNG/O 하/VV/O 았/EP/O 다/EF/O ./SF/O"가 있을 경우에 품사패턴 유사도는 식 (8)과 같이 계산된다.

$$
\begin{aligned}
&\text{품사패턴유사도}(\text{품사}(\text{착한 여자}), \text{품사}(\text{착한 여자 나쁜 여자})) \\
&\quad = \text{품사패턴유사도}(< NP\,JX\,VA\,ETM\,NNG\,JKO\,VV >, \\
&\qquad\qquad < NP\,JX\,VA\,ETM\,NNG\,VA\,ETM\,NNG\,JKO\,NNG >) \\
&\quad = \frac{1}{2\times\left\lfloor \frac{7}{2} \right\rfloor}\begin{pmatrix}(s(NP,NP)+s(VV,NNG)) \\ +(s(JX,JX)+s(JKO,JKO)) \\ +(s(VA,VA)+s(NNG,NNG))\end{pmatrix} \\
&\quad = \frac{1}{6}((1+0)+(1+1)+(1+1)) \\
&\quad = 0.833.
\end{aligned}
\tag{8}
$$

또한, 품사패턴뿐만 아니라, 매 단계의 학습마다 언어정보가 부착된 고정키어구를 수집하여 고정키어구 리스트를 구축하는데, 이는 동일한 문장에 대해 이전의 학습에서 현재의 학습과 동일한 고정키어구 추출 결과를 가질 경우, 선정 대상에서 제외를 시켜, 학습되지 않은 다양한 문장이 선정되도록 하기 위함이다.

4. 실험 및 토의

본 장에서는 각 실험 및 토의에 대해서 기술한다. <표 5>와 <그림 4>에서 보듯이 "실험 3, 4, 5"는 문장신뢰도[1]과 고정키어구 리스트, 품사패턴의 유사도를 조합한 실험의 성능이다. "실험 3"은 17번째 학습에서 지도학습의 성능보다 높은 85.75%를 기록하였다. "실험 4"는 22번째 학습에서 지도학습의 성능보다 높은 86.30%를 기록하였으며, 23번째 학습에서 최고 성능인 86.45%를 기록하였다. "실험 5"는 "실험 3"과 "실험 4"보다 빠른 16번째 학습에서 지도학습의 성능보다 높은 86.26%를 기록하였다.

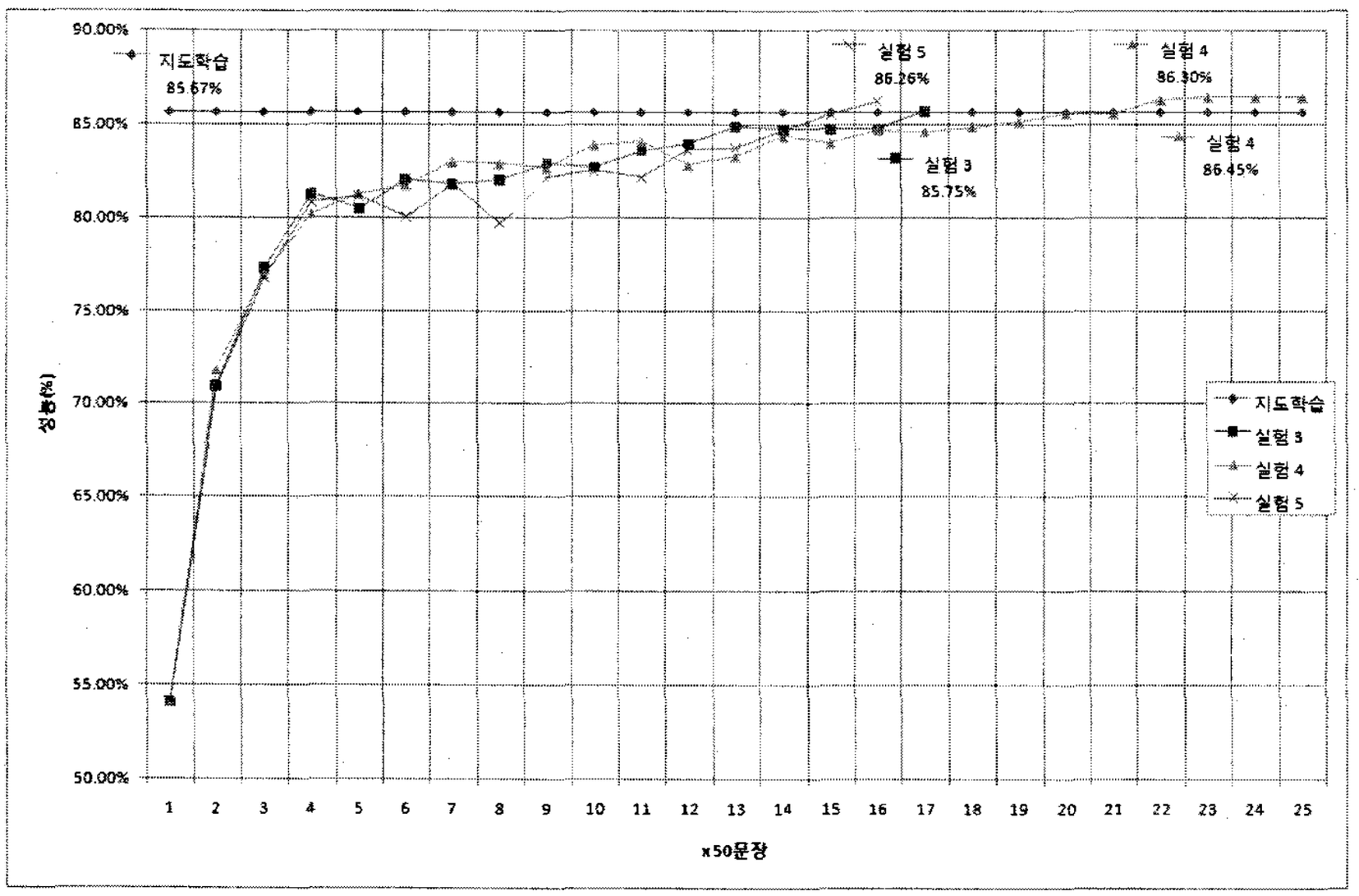
"실험 3, 4, 5"를 통해 고정키어구 리스트를 사용하면 중복된 고정키어구의 학습을 제한한다는 것을 알 수 있다. 특히 품사패턴을 사용하여 다양한 문장을 학습할 경우, 더욱 높은 성능을 보였다. 이는 고정키어구와 함께 주변에 있는 품사패턴도 고정키어구 추출에 중요한 자질이 될 수 있음을 알려준다.

<표 5> "실험 3, 4, 5"의 문장 선정 조건 및 성능(F_{key}) 표

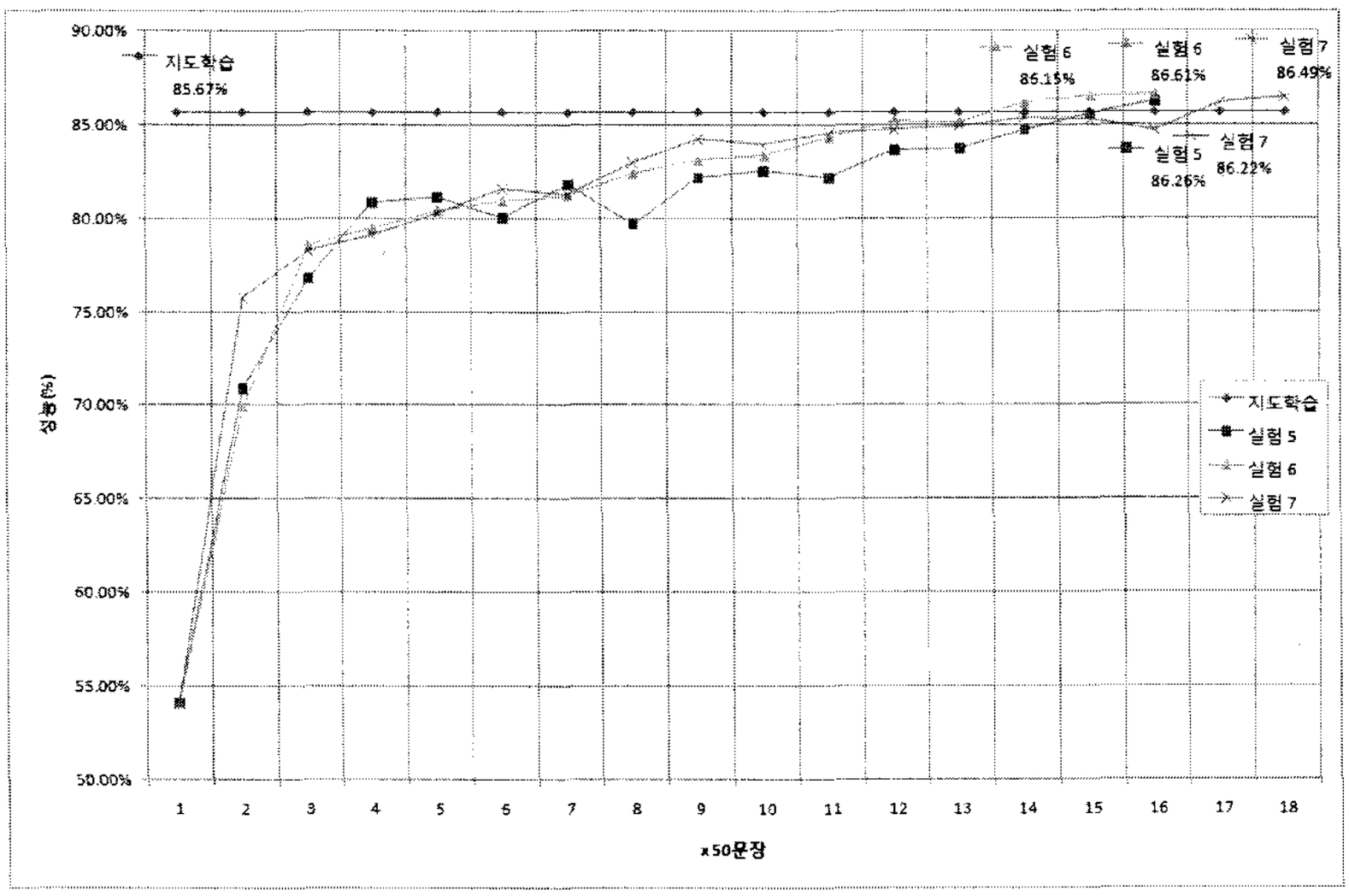
실험 번호	문장 선정 조건	최적 성능		최고 성능	
		단계	성능	단계	성능
실험 3	문장신뢰도[1] 하위 50개, 고정키어구 리스트	17	85.75%	17	85.75%
실험 4	문장신뢰도[1] 하위 50개, 고정키어구 리스트, 품사패턴(size=2) 유사도 0.95이상	22	86.30%	23	86.45%
실험 5	문장신뢰도[1] 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.95이상	16	86.26%	16	86.26%

<표 6> "실험 5, 6, 7"의 문장 선정 조건 및 성능(F_{key}) 표

실험 번호	문장 선정 조건	최적 성능		최고 성능	
		단계	성능	단계	성능
실험 5	문장신뢰도[1] 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.95이상	16	86.26%	16	86.26%
실험 6	문장신뢰도[1] 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.75이상	14	86.15%	16	86.61%
실험 7	문장신뢰도[1] 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.55이상	17	86.22%	18	86.49%



<그림 4> 지도학습 및 "실험 3, 4, 5"의 성능 그래프, F_{key}기준



<그림 5> 지도학습 및 "실험 5, 6, 7"의 성능 그래프, F_{key}기준

Active learning의 최종 목표는 지도학습의 성능에 좀 빨리 접근하는 것이다. 그러므로 <표 5>에서 가장 빨리 지도학습의 성능에 접근한 "실험 5"를 확장하여, 품사패턴의 유사도에 따른 성능 실험의 성능은 <표 6>이며, <그림 5>는 성능을 나타낸 그래프이다. "실험 6"은 "실험 5"보다 빠른 14번째 학습에서 지도학습의 성능보다 높은 86.15%를 기록하였으며, 16번째 학습에서 최고 성능인 86.61%를 기록하였다. "실험 7"은 17번째 학습에서 지도학습의 성능보다 높은 86.22%를 기록하였으며, 18번째 학습에서 최고 성능인 86.49%를 기록하였다.

<표 6>에서 알 수 있듯이 품사패턴의 유사도가 낮으면 다양한 품사패턴이 학습되어 초기에 성능 증가가 빠른 반면, 학습을 반복하면 반복할수록 학습했던 품사패턴과 유사도가 많이 차이나는 문장이 학습되어 학습 말뭉치의 일관성 유지가 힘들어 다른 실험들보다 지도학습의 성능을 넘기는 단계가 늦다. 그렇다면 유사도가 0.95인 "실험 5"에서는 초기의 성능 증가는 느리더라도 어느 실험보다 높은 성능을 보여야 하는데 그렇지 않았다. 그 이유는 다양한 패턴이 추가되지 않아 패턴 증가양이 미약하기 때문이다.

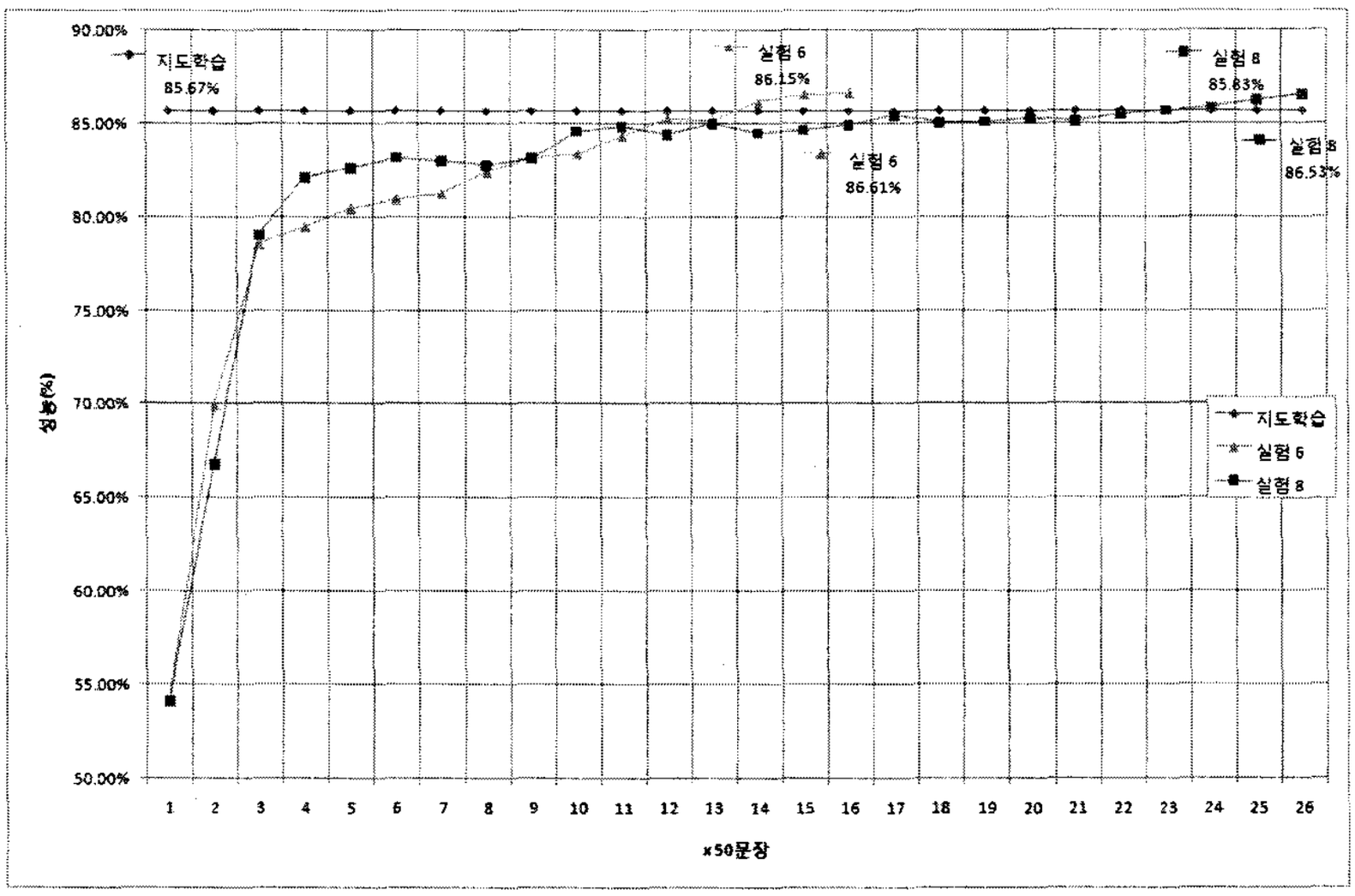
본 논문에서는 "실험 6"에서 사용된 품사패턴을 통한 군집화(품사패턴 유사도 0.75이상)와 문장신뢰도[1](하위 50개 학습)을 통해 다양성과 대표성을 가지는 문장을 선정할 수 있었다.

<표 7>과 <그림 6>은 [5]에서 사용한 문장 CRFs 값(CRFs 정확도)이 성능에 어느 정도 영향을 미치는지 알아본 실험 및 실험의 그래프다. "실험 6, 8"은 모든 조건은 같지만, 학습할 문장 선정 기준인 "문장신뢰도[1] 하위 50개"(실험 6)와 "문장 CRFs 값 하위 50개"(실험 8)만 서로 다르다. 하지만, 전혀 다른 기준임에도 불구하

고 초기의 성능 증가 곡선의 기울기는 비슷하며, 3번째 학습부터는 "실험 8"이 "실험 6"보다 빠른 속도로 성능이 증가하는 것은 "실험 8"이 학습에 좀 더 유용한 문장을 선정한다고 할 수 있다. 하지만, 12번째 학습부터 "실험 8"이 "실험 6"보다 성능이 떨어지는데, 이는 문장신뢰도[1]은 고정키어구가 길면서 많이 등장하는 문장을 우선적으로 학습할 수 있게 하였지만, 문장 CRFs 값은 고정키어구와는 관계없이 모델이 잘 표현할 수 없는 문장만 학습하였기 때문에 고정키어구가 없는 문장이 학습될 수 있기 때문이다.

<표 7> "실험 6, 8"의 문장 선정 조건 및 성능(F_{key}) 표

실험 번호	문장 선정 조건	최적 성능		최고 성능	
		단계	성능	단계	성능
실험 6	문장신뢰도[1] 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.75이상	14	86.15%	16	86.61%
실험 8	문장 CRFs 값 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.75이상	24	85.83%	25	86.53%



<그림 6> 지도학습 및 "실험 6, 8"의 성능 그래프, F_{key}기준

우리는 "실험 8"을 토대로 문장신뢰도[1]을 식 (9)와 같이 변경하였다. 식 (9)는 초기에 성능을 빨리 증가시킬 수 있는 문장을 선정하는 "실험 8"의 문장 CRFs 값과 "실험 6"의 문장신뢰도[1]을 선형 조합한 식으로 α는 문장신뢰도[1]을 우선할 수 있도록 0.75로 정하였다. 이는 학습의 후반에 갈수록 "실험 6"이 "실험 8"보다 좋은 성능을 가지는 특징을 반영한 것이다.

$$KP_j = \langle x_{j,1}, x_{j,2}, x_{j,3}, \cdots, x_{j,h_j} \rangle,$$

$$h_j = |KP_j|,$$

$c =$ 고정키어구 갯수, (9)

$$\text{문장신뢰도}^2 = \alpha \times \frac{1}{c}\sum_{j=1}^{l}\left[\frac{1}{h_j^2}\sum_{r=1}^{h_j} P(x_{j,r})\right] + (1-\alpha)\times(\text{문장}\,CRFs\,\text{값}), \alpha = 0.75.$$

<표 8>은 수정된 문장신뢰도^{2}를 적용한 "실험 9"와 이전의 실험 중에서 좋은 성능을 보였던 "문장신뢰도^1 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.75이상"(실험 6), "문장 CRFs 값 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.75이상"(실험 8)의 결과를 비교한 표이다.

"실험 9"는 "실험 6"과 "실험 8"을 조합한 만큼 "실험 6"보다 두 단계나 앞서 12번째 학습에 지도학습의 성능보다 높은 86.10%를 기록하였다. "실험 6"이 지도학습보다 좋은 성능을 가지는 14번째 학습(86.15%)에서도 "실험 9"가 86.38%를 기록하였지만, 최고 성능은 16번째 학습에서 "실험 6"이 86.61%를 먼저 기록하였으며, 같은 단계에서 "실험 9"는 86.38%를 기록하였다. 18번째 학습에서 "실험 6"의 성능보다 좋은 86.68%를 기록, 최고 성능은 20번째 학습에서 86.80%를 기록하였다.

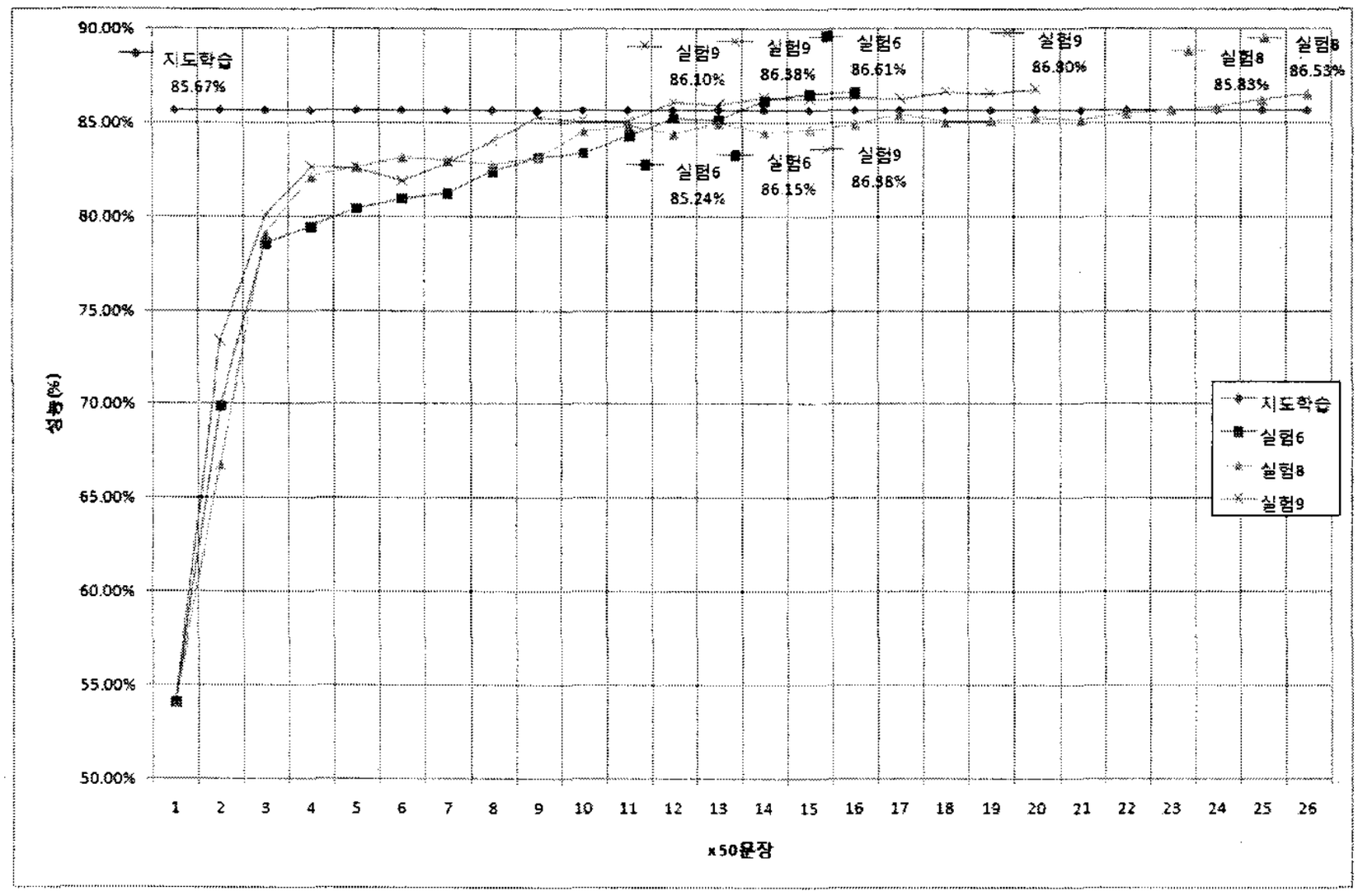
<표 8> "실험 6, 8, 9"의 문장 선정 조건 및 성능(F_{key}) 표

실험 번호	문장 선정 조건	최적 성능		최고 성능	
		단계	성능	단계	성능
실험 6	문장신뢰도^1 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.75이상	14	86.15%	16	86.61%
실험 8	문장 CRFs 값 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.75이상	24	85.83%	26	86.53%
실험 9	문장신뢰도^2 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.75이상	12	86.10%	20	86.80%

<표 9>는 문장신뢰도^{2}를 적용한 "실험 9"와 초기 학습용 말뭉치 크기에 따른 성능을 나타낸 표이며, <그림 8>은 성능을 나타낸 그래프이다. 초기 학습용 말뭉치에 추가된 문장은 실험에 사용되지 않는 동일한 도메인의 말뭉치로부터 임의로 5문장씩 추가하였다.

"실험 10"은 14번째 학습에서 85.95%를 기록하고 학습을 종료하였으며, "실험

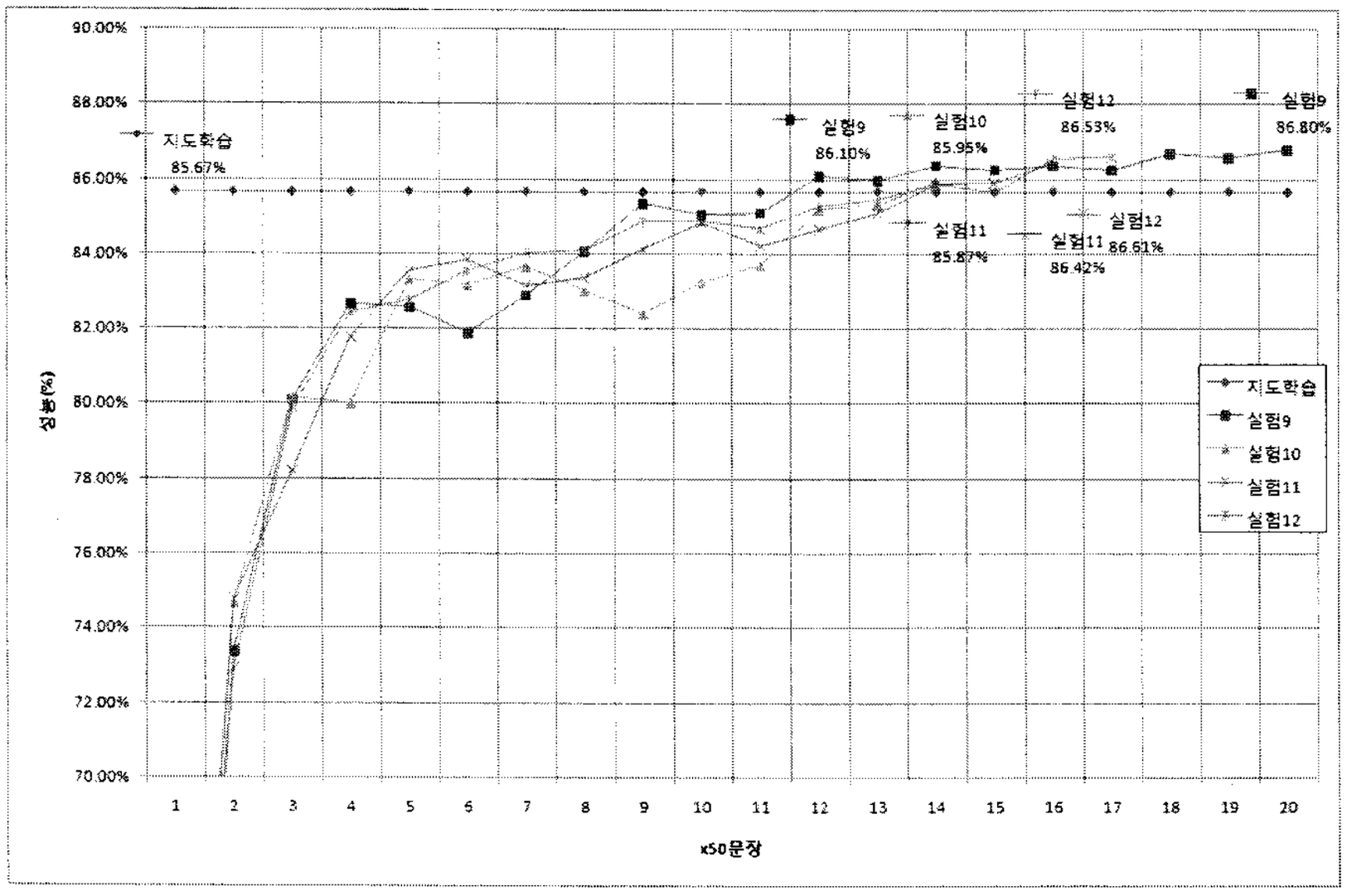
11”은 14번째 학습에서 지도학습 성능보다 좋은 85.87%, 16번째 학습에서 86.42%를 기록하였다. “실험 12”는 16번째 학습에서 지도학습보다 좋은 86.53%를 기록, 17번째 학습에서 최고 성능인 86.53%을 기록하였다. “실험 11, 12”에서 초기 학습용 말뭉치가 증가하더라고 성능 향상에는 영향을 못 미친다는 것을 알 수 있다.



<그림 7> 지도학습 및 “실험 6, 8, 9”의 성능 그래프, F_{key}기준

<표 9> “실험 9, 10, 11, 12”의 문장 선정 조건 및 성능(F_{key}) 표

실험 번호	문장 선정 조건	최적 성능		최고 성능	
		단계	성능	단계	성능
실험 9	문장신뢰도[2] 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.75이상, 초기 학습용 말뭉치 10문장	12	86.10%	20	86.80%
실험 10	문장신뢰도[2] 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.75이상, 초기 학습용 말뭉치 5문장	14	85.95%	14	85.95%
실험 11	문장신뢰도[2] 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.75이상, 초기 학습용 말뭉치 15문장	14	85.87	16	86.42%
실험 12	문장신뢰도[2] 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.75이상, 초기 학습용 말뭉치 20문장	14	85.87%	17	86.61%



<그림 8> 지도학습 및 "실험 6, 8, 9"의 성능 그래프, F_{key}기준

<표 10> "실험 3, 4, 5, 6, 7, 8, 9"의 문장 선정 조건 및 성능(F_{key}) 표

실험 번호	문장 선정 조건	최적 성능		최고 성능	
		단계	성능	단계	성능
-	지도 학습	-	-	-	85.67%
실험 2	문장신뢰도[1] 하위 50개	-	-	10	83.31%
실험 3	문장신뢰도[1] 하위 50개, 고정키어구 리스트	17	85.75%	17	85.75%
실험 4	문장신뢰도[1] 하위 50개, 고정키어구 리스트, 품사패턴(size=2) 유사도 0.95이상	22	86.30%	23	86.45%
실험 5	문장신뢰도[1] 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.95이상	16	86.26%	16	86.26%
실험 6	문장신뢰도[1] 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.75이상	14	86.15%	16	86.61%
실험 7	문장신뢰도[1] 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.55이상	17	86.22%	18	86.49%
실험 8	문장 CRFs 값 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.75이상	24	85.83%	26	86.53%
실험 9	문장신뢰도[2] 하위 50개, 고정키어구 리스트, 품사패턴(size=3) 유사도 0.75이상	12	86.10%	20	86.80%

5. 결론 및 향후 연구

본 논문에서 이루어진 실험을 종합한 <표 10>을 보면 (1) 학습 양을 줄이기 위

해 active learning를 사용하였으며, (2) "실험 2, 3"에서 학습된 고정키어구를 반복하여 학습할 경우, 성능 향상에는 좋지 않다는 것을 알 수 있었다. (3) "실험 3, 4, 5"를 통해, 말뭉치에 있는 문장을 순서대로 학습하는 것 보다는 다양한 품사패턴을 가진 문장을 선택하여 학습하는 것이 적은 양의 학습으로도 지도학습의 성능보다 높은 성능을 얻을 수 있다는 것을 확인할 수 있었다. (4) "실험 3, 5"를 통해 고정키어구뿐만 아니라, 고정키어구 주변의 품사들도 고정키어구 추출에 중요한 자질이 됨을 확인할 수 있었다. (5) "실험 5, 6, 7"을 통해 문장의 품사패턴의 유사도가 너무 높거나 낮아도 성능에 좋지 못하다는 것을 알 수 있었다.

그 결과, 학습용 말뭉치의 크기를 5,010문장(지도학습)에 비해 610문장, 약 88%를 줄이는데 성공하였으며, 성능도 85.67%(지도학습)에서 86.80%(실험 9)로 1.13% 증가하였다. 앞으로는 제안된 방법론을 다양한 분야에서 좀 더 효율적으로 적용할 수 있도록 연구할 예정이다.

참 고 문 헌

[1] S. Tong, *Active Learning Theory and Applications*, Ph.D. Dissertation, Stanford University, 2001.

[2] D. D. Lewis, J. Catlett, "Heterogeneous uncertainty sampling for supervised learning", *Proc. Eleventh International Conference on Machine Learning*, pp. 148-156, 1994.

[3] Y. Freund, H. S. Seung, E. Shamir, "Selective sampling using the query by committee algorithm", *Machine Learning*, Vol. 28, No. 2, pp. 133-168, 1997.

[4] S. Dan, Z. Jie, S. Jian, "Multi-criteria-based active learning for named entity recognition", *Proc. 42nd Annual Meeting on Association for Computational Linguistics*, pp. 589-596, 2004.

[5] 박훈민, *대화 시스템을 위한 CRFs와 Active Learning 기반의 효율적 의미 구조 분석*, 컴퓨터공학과, 서강대학교 대학원, 2006.

[6] J. Lafferty, A. McCallum, F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", *Proc. of 18th International Conference on Machine Learning*, pp. 282-289, 2001.

[7] S. Della Pietra, V. Della Pietra, J. Lafferty, "Inducing features of random fields", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 4, pp. 380-393, 1997.

[8] A. L. Berger, V. J. Della Pietra, S. A. Della Pietra, "A maximum entropy approach to natural language processing", *Computational Linguistics*, Vol. 22, No. 1, pp. 39-71, 1996.

[9] S.-B. Park, Y.-S. Tae, S.-Y. Park, "Self-organizing n-gram model for automatic word spacing", *Proc. 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, pp. 633-640, 2006.

[10] 김현정, 은지현, 장두성, "홈네트워크 제어를 위한 대화관리시스템 설계", *대한음성학회 가을 학술대회 발표논문집*, pp. 109-112, 2006.

접수일자: 2008년 5월 11일
게재결정: 2008년 6월 19일

▶ 이현우(Hyun-Woo Lee)
주소: 641-773 경남 창원시 사림동 9번지 소나무5길 국립창원대학교
소속: 국립창원대학교 컴퓨터공학과 자연어처리연구실
전화: 055) 213-3822 (내선 108)
E-mail: ggamsso@changwon.ac.kr

▶ 차정원(Jeong-Won Cha) : 교신저자
주소: 641-773 경남 창원시 사림동 9번지 소나무5길 국립창원대학교
소속: 국립창원대학교 컴퓨터공학과
전화: 055) 213-3818
E-mail: jcha@changwon.ac.kr