수정된 Active Learning을 이용한 고정키어구 추출

이 현우 O*, 은 지현, 장 두성, 차 정원*
*국립창원대학교 컴퓨터공학과
ggamsso@changwon.ac.kr, jcha@changwon.ac.kr
KT 미래기술연구소
jh06@kt.com, dschang@kt.com

Extraction of Keyphrase using modified Active Learning

Hyun-Woo Lee O*, Ji-Hyun Eun, Du-Seong Jang, Jeong-Won Cha*
*Changwon National University
KT Future Technology Laboratory

요 약

본 연구에서는 Active Learning의 학습과정을 변형하여 학습노력을 줄이고 성능향상을 이루는 방법에 대해서 기술한다. Active Learning을 사용하는 이유는 학습 코퍼스의 량을 줄이면서도 우수한 성능을 얻기 위해서이다. 우리는 학습량을 줄이기 위해서 다양성과 대표성이 높은 학습 데이터를 추가한다. 높은 다양성을 얻기 위해서 기 학습된 코퍼스와 가장 관련이 없는 데이터를 추가하고 높은 대표성을 얻기 위해 예제 군집화를 통해 대표적인 예제를 추가할 수 있도록 하였다. 제안된 방법의 효용성을 검사하기 위해서 고정키어구 추출 문제에 적용하였다. 실험결과를 보면 지도학습을 이용한 실험결과보다 우수하였으며, 학습량을 83%정도 줄일 수 있었다.

1. 서론

고정키어구(keyphrase)는 일반적으로 많이 사용되는 "영화", "드라마"와 같은 하나의 키워드와 "하늘이시여", "착한 여자 나쁜 여자", "TV는 사랑을 싣고"와 같이 하나 이상의 연속된 키워드로 구성된 것을 말한다. 고정키어구를 추출하기 위해서는 미리 "의미 부착 작업"이 완료된 "학습용 말뭉치"를 "기계학습"하는 방법이 주류를 이룬다.

기계학습의 가장 큰 문제점은 의미가 부착된 많은 량의 학습용 말뭉치가 필요하다는 점이다. 학습용 말뭉치의 의미 부착 작업[1]은 막대한 인적, 물적 자원을 필요로 한다. 또한, 작업 자체가 어렵고 까다로우며, 많은 사람이 함께 하는 작업인 관계로 "잘못된 의미"가 많이 부착되는 문제점을 가지고 있다.

본 논문에서는 위와 같은 어려운 의미 부착 작업을 적은 량의 학습용 말뭉치와 Active Learning[1]을 이용하여 보다 효율적으로 해결하고자 한다. Active Learning은 [**오류! 참조 원본을 찾을 수 없습니다.**]과 같이 "초기 학습용 말뭉치"(seed corpus)를 이용하여 모델을 학습하고 이 모델을 이용하여 "의미가 부착되지 않은 말뭉치"(raw corpus)에 의미를 부착한 뒤에 일부를 사용자에게 다시 의미 부착 작업"을 요청한다. 여기서 사용자는 부착된 의미를 수정하게 되며 이를 학습용 말뭉치에 추가하여 다음 학습에 이용하는데, 이러한 과정을 사용자가 만족하는 성능에 도달할 때까지 반복하게 된다.

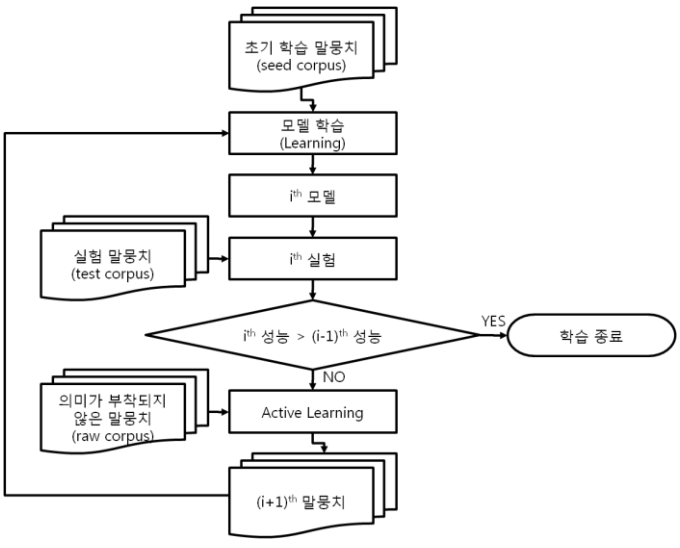

그림 1. Active Learning의 방법

2. 이전연구

Active Learning에서는 사용자에게 "의미 부착 작업"을 요청할 때, 성능을 많이 증가시킬 수 있는 말뭉치를 전달하는 것에 초점을 맞추어 연구가 이루어

[1] 여기서 의미 부착 작업은 품사 부착, 구문 정보 부착 등 자연어처리에서 사용되는 모든 정보를 부착하는 작업을 포함하는 말로 사용한다.

졌다.

[2]는 주어진 문서를 양성예제(positive example)와 음성예제(negative example)로 구별할 때, Naive Bayes 알고리즘으로 양성예제(positive example)에 속할 확률이 0.5인 예제, 즉 가장 구별하기 힘든 예제만을 선정하여 학습하였으며, [3]에서도 [2]와 비슷하게 학습 단계에서 다수의 위원회(committee)를 두어, 위원들 사이에서 동일한 예제에 대하여 각 위원간의 의견이 가장 일치하지 않은 예제만을 선정하여 다음 학습에 이용하는데, 이는 이러한 예제들이 가장 성능을 빨리 올릴 수 있는 의미 있는 예제라고 정의하였기 때문이다.

[4]에서는 개체명을 부착할 때, 의미 있는 예제뿐만 아니라, 개체명 또는 단어 사이의 유사도와 kNN-clustering으로 대표적(Representativeness)이면서 다양한 예제(Diversity)를 선정하여 학습하였다.

[5]에서는 기계학습에 사용된 CRFs의 정확도가 가장 낮은 예제를 선정하여 학습하였다.

3. CRFs를 이용한 고정키어구 추출

본 논문에서는 고정키어구 추출을 품사 부착 문제와 같이 형태소에 바로 태그(tag)를 붙이는 방식으로 해결하였다. 예를 들어 "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 표기된다.

[그림 2]먼저 논문에 사용된 용어에 대해 정의한다. 그림에서 X는 문장을 구성하는 형태소 x들의 벡터이고, Y는 문장을 구성하는 품사 y들의 벡터로 정의한다.

$X = < x_1, x_2, \ldots x_n >$

$Y = < y_1, y_2, \ldots y_n >$

$T = < t_1, t_2, \ldots t_n >$

$t = \{B, I, O\}$

B: 고정키어구의 시작

I: 고정키어구의 중간/끝

O: 고정키어구 아님

그림 2. 용어 정의

3.1. Conditional Random Fields

CRFs는 조건부 확률을 최대로 하는 방향성이 없는 그래프 모델이다[6]. 입력열 $X = x_1 x_2 \ldots x_n$, 상태열 $T = t_1 t_2 \ldots t_n$가 주어졌을 때, CRFs는 조건 확률로 (1)과 같이 정의된다.

$$P(X|T) = \frac{1}{Z_t} \exp\left(\sum_{i=1}^{n}\sum_{k} \lambda_k f_k(x_{i-1}, x_i, t, i)\right). \quad (1)$$

여기서 Z_t는 확률값으로 만들어 주는 정규화 값이고, $f_k(x_{i-1}, x_i, t, i)$ 는 자질 함수이다. 또한 λ_k 는 각 자질에 대한 가중치를 나타낸다. 자질 함수는 현재 시간 i에 대한 관측열 x_i에 대해서 전이의 양상을 측정할 수 있다.

매개변수들은 주어진 입력열과 이에 대응하는 상태열에 대한 조건부 확률을 최대화하는 최대 유사도(maximun likelihood)에 의해서 추정된다. 훈련 집합 $\{(t_l, x_l)\}_{l=1}^{N}$ 에 대해서 다음과 같은 로그 유사도(log-likelihood)(2)를 최대화 하도록 학습한다.

$$\begin{aligned} L(\Lambda) &= \sum_{l} \log P_\Lambda(x_i|t_i) \\ &= \sum_{l}\left(\sum_{i=1}^{n}\sum_{k} \lambda_k f_k(x_{i-1}, x_i, t, i) - \log Z_{t_i}\right). \end{aligned} \quad (2)$$

일반적으로 CRFs는 Improve Iterative Scaling(IIS)나 Generalized Iterative Scaling(GIS)[7]를 사용하여 학습한다. 또한 학습 데이터의 과적합(overfitting) 문제를 해결하기 위해서 가우시안 사전 평활화(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 하였으므로, 본 논문에서도 이와 같은 점을 반영하였다.

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 형태소 "착하"일 경우에 대한 자질이다.

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	$x_i/y_i/x_{i+1}/y_{i+2}$	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	$x_i/x_{i+1}/x_{i+}$	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

3.3.1. 학습 및 실험용 말뭉치

표 2. 학습 및 실험 말뭉치 정보

	초기 학습용 말뭉치	의미가 부착되지 않은 말뭉치	평가용 말뭉치
문장 수	10	5000	300
형태소 수	156	46,036	2,311
평균 형태소 수	15	9	7

본 논문에서는 대화 시스템을 위해 제작된 총 5,310 문장(48,503 형태소)로 구성된 말뭉치[10]을 사용하였으며, 품사부착기의 오류가 실험에 미치는 영향을 최소화하기 위해서 모든 품사 오류를 수정하였다. 초기 학습용 말뭉치 10 문장, 의미가 부착되지 않은 말뭉치 5,000 문장, 평가용 말뭉치 300 문장으로 나누었다. 초기 학습용 말뭉치는 초기 고정키어구 추출 모델을 생성하기 위해 사용하였다.

3.3.2. 평가기준

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.

$$P_{key} = \frac{\text{올바르게 추출한 고정키어구}}{\text{고정키어구 모델이 추출한 고정키어구 수}} \times 100(\%),$$

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

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

3.3.3. 학습할 예제 선정 방법

[5]에서는 문장 전체의 CRFs 신뢰도가 낮은 예제만을 선정하였는데, 본 논문은 고정키어구만의 신뢰도(3)를 계산하여, 고정키어구가 많이 출현하거나, 긴 고정키어구가 출현한 예제 중에 "신뢰도 상위 50개", "신뢰도 상위 25개, 신뢰도 하위 25개", "신뢰도 하위 50개"를 선정하여, 사용자가 의미 부착 작업을 한 뒤, 초기 말뭉치와 합쳐서 학습하는 과정을 반복하였다.

학습을 중단하는 시점은 지도학습(Supervised Learning)으로 학습한 모델의 성능 [표 3]을 능가하거나, 성능 증가가 3회 동안 없을 경우 학습을 중단한다(F_{key} 기준).

고정키어구의 신뢰도

$$= \frac{\sum_{j=1} \frac{\sum_{i=1}(\text{형태소의 신뢰도})_{i,j}}{\text{고정키어구의 형태소 수}}}{\text{고정키어구 수}} \times \log(\text{고정키어구 수}). \quad (3)$$

표 3. 지도학습한 고정키어구 모델의 성능, 단위(%)

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

표 4. "신뢰도 상위 50개", "신뢰도 상위 25개, 신뢰도 하위 25개", "신뢰도 하위 50개"의 실험, 성능이 3회 이상 증가하지 않을 경우 학습 중단, 단위 (%)

실험 번호	1			2			3		
선정 기준	신뢰도 상위 50개			신뢰도 상위 25개 신뢰도 하위 25개			신뢰도 하위 50개		
학습 문장 개수	P_{key}	R_{key}	F_{key}	P_{key}	R_{key}	F_{key}	P_{key}	R_{key}	F_{key}
10	82.51	40.27	54.12	82.51	40.27	54.12	82.51	40.27	54.12
60	87.50	58.49	70.11	75.61	75.41	75.51	87.27	51.47	64.75
110	84.88	66.76	74.74	79.82	71.43	75.39	87.18	63.81	73.69
160	85.81	68.65	76.28	80.47	72.92	76.51	91.12	63.61	74.92
210	86.75	71.00	78.09	80.95	73.51	77.05	91.11	66.67	77.00
260	86.60	71.82	78.52	80.51	77.03	78.73	91.43	69.19	78.77
310	85.35	72.63	78.48	80.00	76.55	78.24	90.66	70.62	79.39
360	**84.69**	**73.44**	**78.66**	81.07	77.36	79.17	90.41	71.35	79.76

410	85.08	72.43	78.25	82.44	78.44	80.39	90.75	71.24	79.82
460	85.08	72.43	78.25	83.15	80.22	81.66	90.51	71.77	80.06
510	-	-	-	83.33	79.73	81.49	89.97	72.31	80.18
560	-	-	-	83.43	80.27	81.82	89.14	73.05	80.30
610	-	-	-	84.08	81.57	82.81	89.07	74.66	81.23
660	-	-	-	85.23	81.08	83.10	89.68	74.93	81.64
710	-	-	-	84.94	81.03	82.94	89.91	76.82	82.85
760	-	-	-	84.94	81.03	82.94	89.44	77.63	83.12
810	-	-	-	84.55	81.57	83.03	89.85	78.49	83.79
860	-	-	-	85.07	81.62	83.31	88.86	79.51	83.93
910	-	-	-	83.94	80.54	82.21	88.86	79.73	84.05
960	-	-	-	85.03	81.35	83.15	88.92	80.27	84.37
1,010	-	-	-	85.11	81.67	83.36	88.66	80.05	84.14
1,060	-	-	-	85.39	81.94	83.63	89.49	80.32	84.66
1,110	-	-	-	85.15	81.94	82.52	89.25	80.81	84.82
1,160	-	-	-	85.39	81.94	83.63	89.88	81.40	85.43
1,210	-	-	-	85.28	82.75	83.99	89.94	82.38	85.99
1,260	-	-	-	85.7	82.48	84.07	**89.97**	**82.43**	**86.04**
1,310	-	-	-	**85.99**	**82.53**	**84.22**	-	-	-
1,360	-	-	-	85.99	82.53	84.22	-	-	-
1,410	-	-	-	85.71	82.26	83.95	-	-	-

[표 4]의 “신뢰도 상위 50개”실험은 초기 실험 말뭉치에 너무 과적합된, 즉 초기 실험 말뭉치와 비슷한 예제들만 선정되는 문제가 있어, 성능이 78.66%에서 학습을 멈추었다. 반면, “신뢰도 하위 50개”실험은 [표 3]를 능가하는 실험 결과가 나왔다. 이는 초기 학습용 말뭉치와는 전혀 다른, 성능을 증가시킬 수 있는 다양한 문장들이 선정되었기 때문이다. 그럼에도 불구하고 지도학습[표 3]보다 성능이 낮은 이유는 성능을 증가시킬 수 있는 문장들을 학습하지 못했기 때문이다.

3.3.4. 대표 학습 후보 선정

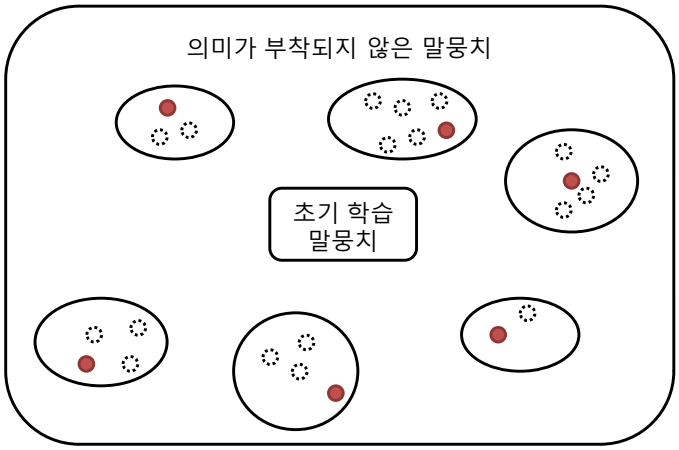

그림 3. 붉은 원은 선정된 예제, 점선 원은 선정되지 않은 예제, 실선으로 그려진 원은 비슷한 예제들의 군집을 나타내고 있다.

본 논문에서는 [그림 3]처럼 다양성과 대표성을 가진 예제를 선정하기 위해, 품사 패턴을 이용한 군집화를 사용한다. 품사 패턴은 고정키어구와 문맥에 있는 품사 패턴이며, 수집된 고정키어구의 품사 패턴이 정확하게 일치하는 군집 중, 고정키어구의 신뢰도가 낮은 한 개의 문장만 선정해서, 매 학습 때마다 일정 개수의 패턴을 학습하도록 한다.

또한, 품사 패턴뿐만 아니라, 매 학습(i번째)에서 의미 부착을 한 고정키어구를 사전으로 제작하여, 다음 학습(i + 1 번째)에서 동일한 추출 결과를 가질 경우, 선정 대상에서 제외를 시켜, 학습한 예제를 선정대상에서 제외시켜, 학습하지 못한 디른 다양한 예제가 선정되도록 하였다.

4. 실험

표 5. 고정키어구 사전 및 품사 패턴을 사용한 실험, (size=2)는 고정키어구 시작과 끝을 기준 앞 뒤 두 개의 품사까지 패턴을 확대, (size=3)은 앞 뒤 세 개의 품사까지 패턴 확대, bold체는 지도학습의 성능을 넘은 시점, 밑줄은 최고 성능, 단위 (%)

실험 번호	4			5			6		
선정 기준	신뢰도 하위 50개 고정키어구 사전			신뢰도 하위 50개 고정키어구 사전 품사 패턴(size=2)			신뢰도 하위 50개 고정키어구 사전 품사 패턴(size=3)		
학습 문장 개수	P_{key}	R_{key}	F_{key}	P_{key}	R_{key}	F_{key}	P_{key}	R_{key}	F_{key}
10	82.51	40.27	54.12	82.51	40.27	54.12	82.51	40.27	54.12
60	82.69	64.20	72.28	83.04	63.51	71.98	82.50	62.43	71.08
110	83.54	74.25	78.62	80.76	75.07	77.81	82.61	77.24	79.83
160	83.24	77.63	80.34	82.07	81.84	81.95	82.86	78.59	80.67
210	81.40	81.40	81.40	83.16	84.28	83.71	82.64	81.30	81.97
260	79.43	82.21	80.80	82.13	83.47	82.80	83.33	84.01	83.67
310	80.73	83.33	82.01	80.98	85.37	83.11	82.45	84.01	83.22
360	81.14	84.86	82.96	81.40	85.37	83.33	82.54	84.32	83.42
410	81.65	85.41	83.49	81.28	85.68	83.42	83.07	84.86	83.96
460	81.96	85.71	83.79	82.01	86.22	84.06	82.46	84.91	83.67
510	82.26	86.25	84.21	81.91	85.68	83.75	83.64	85.68	84.65
560	82.99	87.03	84.96	82.03	85.14	83.55	84.13	85.95	85.03
610	83.64	87.03	85.30	82.17	86.18	84.13	84.17	86.22	85.18
660	83.68	87.53	85.56	82.90	86.72	84.77	84.72	85.41	85.06
710	83.46	87.30	85.34	83.12	86.72	84.88	**85.48**	**85.95**	**85.71**
760	83.72	87.57	85.60	83.68	87.53	85.56	86.50	86.49	86.49
810	**84.72**	**88.62**	**86.63**	**84.38**	**87.80**	**86.06**	87.01	86.99	87.00

[표 5]의 실험번호 4, 5, 6은 [표 4]보다 학습 문장 개수가 약 35%정도 줄어든 것을 확인할 수 있다. 이는 고정키어구 사전을 사용하여 중복된 고정키어구의

학습을 제한했으며, 품사 패턴을 사용하여 다양한 패턴의 문장을 학습한다는 결과이다. 특히 지도학습의 F_{key} 을 능가하는 시점에서는 "실험번호 4"보다 "실험번호 6"의 학습 문장의 수가 100개 작은데 이는 고정키어구와 함께 주변에 있는 품사 패턴도 고정키어구 추출에 중요한 영향을 끼친다고 해석된다.

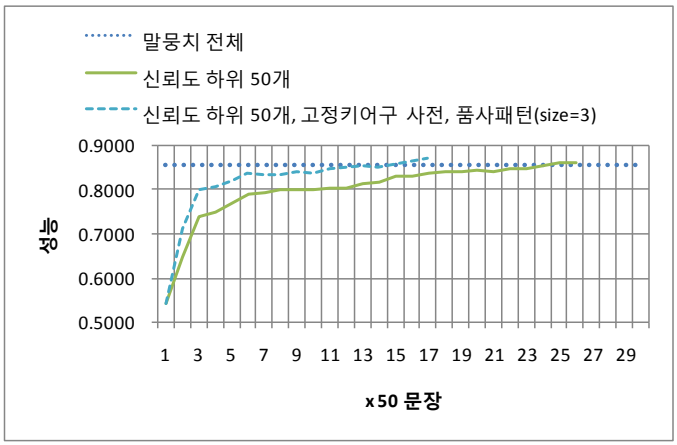

그림 4. "신뢰도 하위 50개", "신뢰도 하위 50개, 고정키어구 사전, 품사패턴(size=3)"의 성능 그래프

5. 결론

본 논문에서는 (1) 학습 량을 줄이기 위해 Active Learning를 사용하였으며, (2) "실험 3, 4"에서 학습된 고정키어구를 반복하여 학습할 경우, 성능 향상에는 좋지 않다는 것을 알 수 있었다. (3) "실험 4, 5, 6"을 통해, 말뭉치에 있는 문장을 순서대로 학습하는 것보다는 다양한 품사패턴을 가진 문장을 선택하여 학습하는 것이 적은 량의 학습으로도 지도학습의 성능보다 높은 성능을 얻을 수 있다는 것을 확인할 수 있었다. (4) "실험 4, 6"을 통해 고정키어구 뿐만 아니라, 고정키어구 주변의 품사들도 고정키어구 추출에 중요한 자질이 됨을 확인할 수 있었다. 그 결과, 학습용 말뭉치의 크기를 5,010문장(지도학습)에 비해 810문장, 약 83%를 줄이는데 성공하였다. 성능도 85.67%에서 87.00%로 1.33% 증가하였다. 앞으로는 제안된 방법론을 다양한 분야에서 좀 더 효율적으로 적용할 수 있도록 연구할 예정이다.

6. 참고문헌

[1] S. Tong, "ACTIVE LEARNING: THEORY AND APPLICATIONS," STANFORD UNIVERSITY, 2001.

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

[3] Y. Freund, H. S. Seung, E. Shamir *et al.*, "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 *et al.*, "Multi-criteria-based active learning for named entity recognition," in Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics, Barcelona, Spain, 2004, pp. 589-596.

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

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

[7] S. Della Pietra, V. Della Pietra, and 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, and 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, and Y.-S. T. S.-Y. Park, "Self-organizing n-gram model for automatic word spacing," in Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL, Sydney, Australia, 2006, pp. 633-640.

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