

# 퓨샷 개체명 인식을 위한 Maximal Marginal Relevance 기반의 라벨 단어 집합 생성

최효림<sup>0</sup>, 황현선, 이창기

강원대학교빅데이터메디컬융합학과 강원대학교컴퓨터공학과  
rimille@kangwon.ac.kr, hhs4322@kangwon.ac.kr, [leeck@kangwon.ac.kr](mailto:leeck@kangwon.ac.kr)

## Generating Label Word Set based on Maximal Marginal Relevance for Few-shot Name Entity Recognition

HyoRim Choi<sup>0</sup>, Hyunsun Hwang, Changki Lee  
Department of Big Data Medical Convergence, Kangwon National University

### 요약

최근 다양한 거대 언어모델(Large Language Model)들이 개발되면서 프롬프트 엔지니어링의 대한 다양한 연구가 진행되고 있다. 본 논문에서는 퓨샷 학습 환경에서 개체명 인식의 성능을 높이기 위해서 제안된 템플릿이 필요 없는 프롬프트 튜닝(Template-free Prompt Tuning) 방법을 이용하고, 이 방법에서 사용된 라벨 단어 집합 생성 방법에 Maximal Marginal Relevance 알고리즘을 적용하여 해당 개체명에 대해 보다 다양하고 구체적인 라벨 단어 집합을 생성하도록 개선하였다. 실험 결과, 'LOC' 타입을 제외한 나머지 개체명 타입에서 'PER' 타입은 0.60%p, 'ORG' 타입은 4.98%p, 'MISC' 타입은 1.38%p 성능이 향상되었고, 전체 개체명 인식 성능은 1.26%p 향상되었다. 이를 통해 본 논문에서 제안한 라벨 단어 집합 생성 기법이 개체명 인식 성능 향상에 도움이 됨을 보였다.

주제어: 개체명 인식, Few-shot Learning, Template-free Prompt Tuning, Label Word Set, Maximal Marginal Relevance

### 1. 서론

최근 다양한 거대 언어모델(Large Language Model)들이 개발되면서 사용자의 의도를 잘 파악하고 더 좋은 답변을 주기 위한 최적의 프롬프트(Prompt)를 찾는 작업인 프롬프트 엔지니어링의 대한 다양한 연구가 진행되고 있다[1,2]. 템플릿 기반 프롬프트 튜닝(Template-based Prompt Tuning)은 문장 수준의 자연어처리 태스크에 적용되는 프롬프트 엔지니어링의 한 방법으로 사전에 정의된 문장 구조의 템플릿(Template)을 입력 문장에 추가하여 사전 학습된 언어 모델이 라벨 단어(Label Word)를 생성하도록 하여 분류 작업 등을 수행하는 방식으로 문장 수준의 자연어처리 태스크에서 성능이 향상됨을 보였으며, 기존의 미세 조정(Fine-tuning) 방식과 다르게 추가적인 파라미터 없이 사전 학습된 언어 모델을 그대로 사용하여 적은 학습데이터로도 높은 성능을 낼 수 있어 퓨샷 학습(Few-Shot Learning) 환경에 유리하다는 장점이 있다[2]. 그러나, 템플릿 기반 프롬프트 튜닝을 개체명 인식과 같은 토큰 수준의 자연어 처리 태스크에 적용하려면 모든 가능한 개체명에 대해 템플릿을 열거하는 과정이 필요하기 때문에 시간이 오래 걸린다는 단점이 존재한다. 이러한 단점을 극복하기 위해 [3]에서는 개체명 인식을 위해 템플릿이 필요 없는 프롬프트 튜닝(Template-free Prompt Tuning) 방식을 제안하고, 라벨 단어 집합(Label Word Set)을 생성하기 위해 단어의 빈도수와 언어 모델

의 확률을 이용하는 방법을 제안하였다.

본 논문에서는 퓨샷 학습 환경에서 개체명 인식의 성능을 높이기 위해 [3]에서 제안한 템플릿이 필요 없는 프롬프트 튜닝 방법을 이용하고, 이 방법에서 사용된 라벨 단어 집합 생성 방법을 개선하였다. 이를 위해 기존의 라벨 단어 집합 생성 방법에 Maximal Marginal Relevance(MMR) 알고리즘을 적용하여 해당 개체명에 대해 보다 다양하고 구체적인 라벨 단어 집합을 생성하였다. 제안한 라벨 단어 집합 생성법을 사용하여 실험한 결과 기존 라벨 단어 집합을 사용했을 때보다 성능이 향상됨을 확인하였다.

### 2. 관련 연구

#### 2.1 개체명 인식

개체명 인식에 관한 연구는 정보 추출 방법의 개발을 장려하기 위해 주최된 Message Understanding Conference (MUC)로부터 논의되기 시작했다[4]. 연구 초기에는 연구의 대상이 거의 영어에 치중되었지만 최근에는 다양한 언어에 대한 연구가 활발히 진행되고 있다[5]. 개체명 인식 연구에는 주로 BIO 태그를 활용한 Sequence labeling 기법이 정형화 되어있고 최근에는 RNN, LSTM, BERT 등과 같은 딥러닝 기반 모델이 주로 사용되고 있으며[6], 최근에는 적은 학습 데이터로도 높은 성능을 보일 수 있는 퓨샷 학습 기법 등이 연구되고 있다[3].

## 2.2 프롬프트 엔지니어링

최근 GPT 모델과 같은 거대 언어모델을 이용하여 특정 작업을 템플릿과 라벨 단어를 사용하여 새로운 작업으로 재구성하는 프롬프트 엔지니어링 기법들이 활발히 연구되고 있으며[7], 프롬프트 엔지니어링에 중요한 영향을 미치는 템플릿 구성 방식에 대한 다양한 연구가 진행되었다. [8,9]는 텍스트 분류 작업을 수행하기 위해서 수동으로 정의된 템플릿을 사용하였고, [10]은 템플릿을 자동으로 검색하는 마이닝 접근 방식을 연구하였으며, [11]은 기율기 기반 접근 방식으로 최적의 이산적 템플릿을 검색하는 방법을 제안하였고, [12]는 사전 훈련된 T5모델을 이용하여 템플릿을 생성하는 기법을 제안하였다. [3]은 개체명 인식과 같은 토큰 수준의 자연어처리 태스크에 적용하기 위한 템플릿이 필요 없는 프롬프트 튜닝(Template-free Prompt Tuning) 방법을 제안하였다. 본 논문에서는 [3]에서 제안한 템플릿이 필요 없는 프롬프트 튜닝 방식의 라벨 단어 집합 생성 방법을 개선하기 위해 Maximal Marginal Relevance(MMR) 알고리즘을 적용하였다.

## 2.3 Maximal Marginal Relevance

[13]에서는 문서 검색과 문서 요약을 위해 Maximal Marginal Relevance(MMR)을 제안하여 중복성을 최소화하면서 관련성을 최대화하는 문서 또는 항목을 선택하는 방법을 제안하였다. MMR 알고리즘의 수식은 다음과 같다.

$$\arg \max_{S_i \in D \setminus R} [\alpha Sim_1(S_i, Q) - (1 - \alpha) \max_{S_j \in R} Sim_2(S_i, S_j)] \quad (1)$$

$D$ 는 문서를 이루고 있는 문장들의 전체 집합을 의미하고  $S$ 는 문서 내의 문장,  $R$ 은  $D$ 의 부분집합으로 이미 선택된 문장들을 의미한다.  $Sim_1$ 과  $Sim_2$ 는 유사도 함수를 의미하며 서로 같은 함수를 사용하거나 다른 함수를 사용할 수 있다.

## 3. 퓨샷 개체명 인식

본 논문에서는 학습 데이터가 부족한 환경의 개체명 인식을 위해 [3]에서 제안한 템플릿이 필요 없는 프롬프트 튜닝 방법을 이용하고, 이 방법에서 사용된 라벨 단어 집합 생성 방법에 MMR 알고리즘을 적용하여 해당 개체명에 대해 보다 다양하고 구체적인 라벨 단어 집합을 생성하여 개체명 인식의 성능을 높인다.

### 3.1 개체명 인식을 위한 템플릿이 필요 없는 프롬프트 튜닝

개체명 인식을 위한 템플릿이 필요 없는 프롬프트 튜닝은 그림 1-b와 같이 개체명의 타입과 각 개체명 타입에 대한 라벨 단어가 쌍으로 맵핑 되어진 라벨 단어 집합을 생성하여, 입력 문장 내의 개체명에 해당하는 단어는 해당 개체명에 대한 라벨 단어가 출력되고, 개체명

이 아닌 단어는 입력된 원본 단어 그대로를 출력하도록 사전 학습된 언어 모델을 미세조정(Fine-tuning)하는 기법이다[3]. 예를 들어, 모델에 “Obama was born in America”라는 문장이 입력되면 언어모델은 “PER” 라벨에 해당하는 “Obama”라는 단어의 자리에 “John”이라는 라벨 단어를 예측하고, 개체명에 해당하지 않는 단어인 “was”는 원본 단어 “was”를 그대로 예측한다. 이때 사용되는 라벨 단어는 불연속 라벨 단어(Discrete Label Word)와 벡터 형식의 가상 라벨 단어(Virtual Label Word) 두 종류가 있는데, 불연속 라벨 단어는 단어의 예측 확률 분포와 문서 내 상위 빈도 단어 등을 이용하여 구하고, 가상 라벨 단어는 각 개체명 타입의 상위 빈도 단어들의 평균 벡터를 사용한다. [3]에서는 가상 라벨 단어를 사용한 경우의 개체명 인식 성능이 더 높았기 때문에, 본 논문에서도 가상 라벨 단어를 사용한다.

그러나 [3]의 방식을 통해 구축된 가상 라벨 단어들은 개체명 타입의 상위 빈도 단어들의 평균 벡터를 사용하였기 때문에 평균 벡터 계산시에 추상적인 의미의 단어들이 사용되고, 또한 서로 유사한 의미의 단어들이 사용되어 다양한 의미를 포함한 가상 라벨 단어를 구할 수 없다는 한계가 있다. 이러한 문제를 해결하기 위해, 본 논문에서는 가상 라벨 단어 생성에 MMR 알고리즘을 적용하여 좀더 구체적이고 다양한 단어들로 구성된 평균 벡터를 생성할 수 있도록 한다.

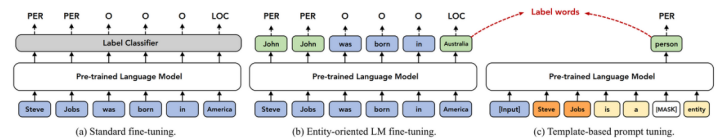


그림 1. Template-free Prompt Tuning 예시[3]

### 3.2 MMR을 적용한 라벨 단어 집합 생성

본 논문에서는 가상 라벨 단어 생성을 위해 상위 빈도 단어들을 선택할 때에 이미 선택된 상위 빈도의 단어와 새로 선택될 후보 단어 사이의 중복성을 줄이고 다양성을 높이기 위해 MMR 알고리즘을 적용하였고, 그 수식은 다음과 같다.

$$\arg \max_{W_i \in D \setminus R} [\alpha \cdot score(W_i) - (1 - \alpha) \max_{W_j \in R} Sim(W_i, W_j)] \quad (2)$$

위 식에서  $\alpha$ 는 중복성과 다양성을 조절하는 상수이고,  $D$ 는 타겟 도메인의 텍스트 데이터(Unlabeled Data)이고,  $R$ 은 가상 라벨 단어 생성을 위해 선택된 상위 빈도의 단어 집합이고,  $score(W_i)$ 는 후보 단어가 텍스트 데이터 내에서 나타나는 빈도수의 log값이며,  $Sim(W_i, W_j)$ 는 선택된 상위 빈도의 단어와 후보 단어 간의 코사인 유사도(Cosine Similarity)를 나타낸다. 코사인 유사도는 python 라이브러리인 ‘Gensim’의 ‘Word2Vec’을 이용하여 구한다[15,16].

표 1은 제안한 기법을 통해 구축한 라벨 단어 집합과 기존 라벨 단어 집합 중 ‘ORG’ 개체명의 라벨 단어 집합을 비교한 것이다. 기존 라벨 단어 집합의 단어들은

추상적이고 유사한 의미의 단어들(Corp와 Inc 등)로 구성되어 있으나, 제안한 기법을 통해 얻은 라벨 단어 집합의 단어들은 프랑스의 도시를 나타내는 'Marseille' 나 영국의 카운티를 나타내는 'Essex', 유럽 연합을 나타내는 'EU' 와 같이 구체적이고 다양한 의미의 단어로 구성되어 있음을 확인할 수 있다.

표 1. 기존 라벨 단어 집합과 MMR을 적용한 라벨 단어 집합 비교

'ORG' 라벨 단어 집합	
기존	"Corp", "Inc", "Commission", "Union", "Bank", "Party"
제안기법	'Corp', 'Los', 'Marseille', 'Union', 'EU', 'Essex'

#### 4. 실험 및 성능

본 논문에서는 [3]과 동일하게 개체명 인식 데이터인 'CoNLL2003' 을 사용하여 퓨샷 학습 환경에서 실험을 진행하였다[3,14].  $\alpha$ 값은 0.5, 0.1, 0.05,0.01,0.001를 사용하였고, 가상 라벨 단어를 구하기 위한 상위 빈도 단어의 개수(Top k)는 [3]과 동일한 6을 사용하였다.  $\alpha$  에 따른 라벨 단어 집합은 부록 A에 비교해 두었다.

표 2는 다양성을 조절하는  $\alpha$ 값에 따른 퓨샷 학습 실험 결과이다. 다양성을 조절하는  $\alpha$ 값을 0.1로 설정하였을 때 가장 좋은 성능을 보임을 알 수 있다.

표 2.  $\alpha$ 에 따른 퓨샷 학습 실험 결과 (F1 score)

	기존	0.5	0.2	0.1	0.05	0.01	0.001
5-shot	51.3 2	55.0 1	54.6 3	<b>55.3</b> 8	52.7 8	48.6 4	54.9 3
10-shot	66.8 6	67.6 3	68.8 6	<b>70.4</b> 6	66.0 4	63.2 3	63.3 5
20-shot	71.2 3	<b>72.9</b> 0	71.4 0	71.5 4	72.8 5	68.2 4	67.7 1
50-shot	74.8 0	74.0 4	74.5 3	<b>76.4</b> 6	76.1	73.9 0	74.0 4

표 3은 가상 라벨 단어를 구하기 위한 상위 빈도 단어의 개수(Top k)에 대한 실험 결과이다. 이때  $\alpha$ 는 0.1, 퓨샷은 50-shot으로 설정하여 진행하였다. 실험 결과, 라벨 단어 집합의 단어 수(Top k)가 변할 경우에도 제안한 기법을 통해 얻은 라벨 단어 집합이 기존 방법으로 구한 라벨 단어 집합보다 높은 성능을 보였다. 각 Top k에 따른 라벨 단어 집합의 구체적인 내용은 부록 B에 비교해 두었다.

표 3. 가상 라벨 단어를 구하기 위한 상위 빈도 단어의 개수(Top k)에 대한 실험

Top k	4	6	10	15	20
Method					
기존	<b>75.26</b>	74.80	73.11	74.37	74.00

제안기법	76.33 (+1.07)	76.46 (+1.66)	75.82 (+2.71)	<b>76.52</b> (+2.15)	76.25 (+2.25)
------	------------------	------------------	------------------	-------------------------	------------------

표 4는  $\alpha = 0.1$ , Top k = 15, 퓨샷은 50-shot일 경우의 개체명 타입별 F1 score를 나타낸다. 'LOC' 타입을 제외하고는 각 타입의 F1 score가 향상되었다. 특히 가장 변화된 단어가 많았던 'ORG' 타입이 4.98%p로 가장 큰 향상 폭을 보였다. 이를 통해 MMR 알고리즘을 적용한 라벨 단어 집합이 개체명 인식 성능 향상에 도움이 되는 것을 알 수 있다.

표 4. 개체명 타입 별 성능 비교 (F1 score)

	기존	제안기법
LOC	<b>75.24</b>	74.94 (-0.3)
MISC	63.09	<b>64.47</b> (+1.38)
ORG	61.85	<b>66.83</b> (+4.98)
PER	92.30	<b>92.90</b> (+0.60)

#### 5. 결론

본 논문에서는 퓨샷 학습 환경에서 개체명 인식의 성능을 높이기 위해 [3]에서 제안한 템플릿이 필요 없는 프롬프트 튜닝 방법을 이용하고, 이 방법에서 사용된 라벨 단어 집합 생성 방법에 MMR 알고리즘을 적용하여 각 개체명 타입에 대해 보다 다양하고 구체적인 라벨 단어 집합을 생성하였다. 실험 결과, 'LOC' 타입을 제외한 나머지 개체명 타입에서 'PER' 타입은 0.60%p, 'ORG' 타입은 4.98%p, 'MISC' 타입은 1.38%p 성능이 향상되었고, 전체 개체명 인식 성능은 1.26%p 향상되었다. 이를 통해 본 논문에서 제안한 라벨 단어 집합 생성 기법이 개체명 인식 성능 향상에 도움이 됨을 보였다.

향후 연구로는 개체명 인식에 적용할 수 있는 다양한 퓨샷 학습 기법에 대한 연구를 진행할 예정이다.

#### 감사의 글

본 논문은 연구재단 과제(NRF2021R1F1A106440311, 딥러닝 기반의 한국어 텍스트 스타일 변환 기술 연구)의 지원을 받아 수행된 연구임

#### 참고문헌

[1] Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." Advances in Neural Information Processing Systems 35 (2022): 24824-24837.

[2] Zhang, Tianyi, et al. "TempLM: Distilling Language Models into Template-Based Generators." arXiv preprint arXiv:2205.11055 (2022).

[3] Ma, Ruotian, et al. "Template-free prompt tuning for few-shot NER." arXiv preprint arXiv:2109.13532 (2021).

[4] Borthwick, Andrew, et al. "NYU: Description of the MENE named entity system as used in MUC-7." Seventh

- Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. 1998.
- [5] Han, Aaron L -F., Derek F. Wong, and Lidia S. Chao. "Chinese named entity recognition with conditional random fields in the light of Chinese characteristics." Language Processing and Intelligent Information Systems: 20th International Conference, IIS 2013, Warsaw, Poland, June 17-18, 2013. Proceedings. Springer Berlin Heidelberg, 2013.
- [6] 이창기, et al. "딥 러닝을 이용한 개체명 인식." 한국정보과학회 학술발표논문집 (2014): 423-425.
- [7] Radford, Alec, et al. "Language models are unsupervised multitask learners." OpenAI blog 1.8 (2019): 9.
- [8] Schick, Timo, and Hinrich Schütze. "Exploiting cloze questions for few shot text classification and natural language inference." arXiv preprint arXiv:2001.07676 (2020).
- [9] Schick, Timo, and Hinrich Schütze. "It's not just size that matters: Small language models are also few-shot learners." arXiv preprint arXiv:2009.07118 (2020).
- [10] Jiang, Zhengbao, et al. "How can we know what language models know?." Transactions of the Association for Computational Linguistics 8 (2020): 423-438.
- [11] Shin, Taylor, et al. "Autoprompt: Eliciting knowledge from language models with automatically generated prompts." arXiv preprint arXiv:2010.15980 (2020).
- [12] Gao, Tianyu, Adam Fisch, and Danqi Chen. "Making pre-trained language models better few-shot learners." arXiv preprint arXiv:2012.15723 (2020).
- [13] Carbonell, Jaime, and Jade Goldstein. "The use of MMR, diversity-based reranking for reordering documents and producing summaries." Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. 1998.
- [14] Sang, Erik F., and Fien De Meulder. "Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition." arXiv preprint cs/0306050 (2003).
- [15] Srinivasa-Desikan B. Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras. Packt Publishing Ltd; 2018.
- [16] Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems 26 (2013).

## 부록

A.  $\alpha$  값에 따른 라벨 단어 집합

	기준
PER	"Michael", "John", "David", "Thomas", "Martin", "Paul"
ORG	"Corp", "Commission", "Inc", "Union", "Party", "Bank"
LOC	"England", "Germany", "Australia", "France", "Russia", "Italy"
MISC	"Palestinians", "Russian", "Chinese", "Russians", "English", "Olympic"

	$\alpha = 0.5$
PER	"Michael", "John", "David", "Martin", "Thomas", "Kong"
ORG	"Corp", "Commission", "Inc", "Union", "Party", "Bank"
LOC	"England", "Germany", "Australia", "France", "Russia", "Italy"
MISC	"Palestinians", "Russian", "Chinese", "Dutch", "English", "Olympic"

	$\alpha = 0.2$
PER	"Michael", "Kong", "Clinton", "Johnson", "Hong", "Le", "Ahmed"
ORG	"Corp", "Union", "Commission", "Party", "Los", "EU", "Marseille"
LOC	"England", "Germany", "Australia", "France", "Russia", "Iraq", "China"
MISC	"Palestinians", "Russian", "Chinese", "English", "Dutch", "Olympic", "EASTERN"

	$\alpha = 0.1$
PER	'Michael', 'Clinton', 'Kong', 'Johnson', 'Hong', 'Le'
ORG	'Corp', 'Los', 'Marseille', 'Union', 'EU', 'Essex'
LOC	'England', 'Iraq', 'Germany', 'Russia', 'Australia', 'China'
MISC	'Palestinians', 'English', 'Russian', 'EASTERN', 'Major', 'CENTRAL'

	$\alpha = 0.05$
PER	'Michael', 'Clinton', 'Kong', 'Hong', 'Hasina', 'Johnson'
ORG	'Corp', 'Essex', 'Marseille', 'Los', 'Hamburg', 'Sussex'

LOC	'England', 'Iraq', 'Moscow', 'Jordan', 'Indonesia', 'Russia'
MISC	'Palestinians', 'EASTERN', 'Major', 'English', 'CENTRAL', 'WESTERN'

	$\alpha = 0.01$
PER	'Michael', 'Hasina', 'Keter', 'Bahadur', 'Azad', 'Bengali'
ORG	'Corp', 'Strasbourg', 'Pauli', 'Essex', 'Hamburg', 'Hassania'
LOC	'England', 'Mahala', 'PACIFIC', 'DAKOTA', 'KS', 'RED'
MISC	'Palestinians', 'Major', 'EASTERN', 'CENTRAL', 'ERA', 'NATIONAL'

	$\alpha = 0.001$
PER	'Michael', 'Hasina', 'Keter', 'Bengali', 'Azad', 'Bahadur'
ORG	'Corp', 'Strasbourg', 'Hassania', 'Pauli', 'Krylya', 'Zizkov'
LOC	'England', 'Mahala', 'PACIFIC', 'DAKOTA', 'KS', 'RED'
MISC	'Palestinians', 'ERA', 'Major', 'EASTERN', 'CENTRAL', 'NATIONAL'

B. Top k에 따른 라벨 단어 집합( $\alpha = 0.1$ )

	$k = 4$ (기준)
PER	"Michael", "John", "David", "Thomas"
ORG	"Corp", "Inc", "Commission", "Union"
LOC	"England", "Germany", "Australia", "France"
MISC	"Palestinians", "Russian", "Chinese", "Dutch"

	$k = 4$ (제안기법)
PER	"Michael", "Clinton", "Kong", "Johnson"
ORG	"Corp", "Los", "Marseille", "Union"
LOC	"England", "Iraq", "Germany", "Russia"
MISC	"Palestinians", "English", "Russian", "EASTERN"

	$k = 10$ (기준)
PER	"Michael", "John", "David", "Thomas", "Martin", "Paul", "Johnson", "Kong", "Clinton", "Robert"
ORG	"Corp", "Inc", "Commission", "Union", "Bank", "Party", "Co", "Ltd", "Council", "House"

## 제35회 한글 및 한국어 정보처리 학술대회 논문집 (2023년)

LOC	"England", "Germany", "Australia", "France", "Russia", "Italy", "India", "Belgium", "China", "London"
MISC	"Palestinians", "Russian", "Chinese", "Dutch", "Russians", "English", "Palestinian", "Olympic", "Republican", "Turkish"

	<b>k = 10 (제안기법)</b>
PER	"Michael", "Clinton", "Kong", "Johnson", "Hong", "Le", "Williams", "Ahmed", "Jones", "La"
ORG	"Corp", "Los", "Marseille", "Union", "EU", "Essex", "Party", "Sussex", "Hamburg", "Utrecht"
LOC	"England", "Iraq", "Germany", "Russia", "Australia", "China", "Moscow", "France", "Israel", "Indonesia"
MISC	"Palestinians", "English", "Russian", "EASTERN", "Major", "CENTRAL", "Chinese", "WESTERN", "Dutch", "DUTCH"

	<b>k = 15 (기존)</b>
PER	"Michael", "John", "David", "Thomas", "Martin", "Paul", "Johnson", "Kong", "Clinton", "Robert", "Hong", "Scott", "Peter", "Ahmed", "Tom"
ORG	"Corp", "Inc", "Commission", "Union", "Bank", "Party", "Co", "Ltd", "Council", "House", "Association", "Department", "EU", "Ministry", "Angeles"
LOC	"England", "Germany", "Australia", "France", "Russia", "Italy", "India", "Belgium", "China", "London", "Netherlands", "Britain", "Spain", "Iraq", "Sweden"
MISC	"Palestinians", "Russian", "Chinese", "Dutch", "Russians", "English", "Palestinian", "Olympic", "Republican", "Turkish", "Indian", "Wimbledon", "EASTERN", "CENTRAL", "WESTERN"

	<b>k = 15 (제안기법)</b>
PER	"Michael", "Clinton", "Kong", "Johnson", "Hong", "Le", "Williams", "Ahmed", "Jones", "La", "Smith", "Khan", "Davis", "De", "States"
ORG	"Corp", "Los", "Marseille", "Union", "EU", "Essex", "Party", "Sussex", "Hamburg", "Utrecht", "Milan", "Yankees", "House", "Barcelona", "Commission"
LOC	"England", "Iraq", "Germany", "Russia", "Australia", "China", "Moscow", "France", "Israel", "Indonesia", "Beijing", "Belgium", "Jordan", "India", "Ukraine"
MISC	"Palestinians", "English", "Russian", "EASTERN", "Major", "CENTRAL", "Chinese", "WESTERN", "Dutch", "DUTCH", "Olympic", "NATIONAL", "ENGLISH", "Indian", "AMERICAN"

	<b>k = 20 (기존)</b>
PER	"Michael", "John", "David", "Thomas", "Martin", "Paul", "Johnson", "Kong", "Clinton", "Robert", "Hong", "Scott", "Peter", "Ahmed", "Tom", "Le", "Jones", "Smith", "Ian", "Williams"

<b>ORG</b>	"Corp", "Inc", "Commission", "Union", "Bank", "Party", "Co", "Ltd", "Council", "House", "Association", "Department", "EU", "Ministry", "Angeles", "Los", "Marseille", "Court", "Senate", "Yankees"
<b>LOC</b>	England", "Germany", "Australia", "France", "Russia", "Italy", "India", "Belgium", "China", "London", "Netherlands", "Britain", "Spain", "Iraq", "Sweden", "Pakistan", "Canada", "Israel", "Europe", "Brazil"
<b>MISC</b>	"Palestinians", "Russian", "Chinese", "Dutch", "Russians", "English", "Palestinian", "Olympic", "Republican", "Turkish", "Indian", "Wimbledon", "EASTERN", "CENTRAL", "WESTERN", "Major", "Serbs", "DUTCH", "Jews", "ENGLISH"

	<b>k = 20 (제안기법)</b>
<b>PER</b>	"Michael", "Clinton", "Kong", "Johnson", "Hong", "Le", "Williams", "Ahmed", "Jones", "La", "Smith", "Khan", "Davis", "De", "States", "York", "Africa", "Mohammad", "van", "Mo"
<b>ORG</b>	"Corp", "Los", "Marseille", "Union", "EU", "Essex", "Party", "Sussex", "Hamburg", "Utrecht", "Milan", "Yankees", "House", "Barcelona", "Commission", "Angeles", "Leeds", "Madrid", "Philadelphia", "Warwickshire"
<b>LOC</b>	"England", "Iraq", "Germany", "Russia", "Australia", "China", "Moscow", "France", "Israel", "Indonesia", "Beijing", "Belgium", "Jordan", "India", "Ukraine", "Netherlands", "Turkey", "Iran", "Italy", "Egypt"
<b>MISC</b>	"Palestinians", "English", "Russian", "EASTERN", "Major", "CENTRAL", "Chinese", "WESTERN", "Dutch", "DUTCH", "Olympic", "NATIONAL", "ENGLISH", "Indian", "AMERICAN", "Celsius", "Republican", "Wimbledon", "Irish", "Scottish"