SVM-based Protein Name Recognition using Edit-Distance Features Boosted by Virtual Examples

가상 예제와 Edit-distance 자질을 이용한 SVM 기반의 단백질명 인식

Eunji Yi[1*], Gary Geunbae Lee[2], Soo-Jun Park[3]

[1,2] Department of Computer Science and Engineering, POSTECH. Pohang, Korea
[3] Bioinformatics Research Team, Computer and Software Research Lab, ETRI. Daejeon, Korea

E-mail: {imew*, gblee}@nlp.postech.ac.kr, psj@etri.re.kr

Abstract

In this paper, we propose solutions to resolve the problem of many spelling variants and the problem of lack of annotated corpus for training, which are two among the main difficulties in named entity recognition in biomedical domain. To resolve the problem of spelling variants, we propose a use of edit-distance as a feature for SVM. And we propose a use of virtual examples to automatically expand the annotated corpus to resolve the lack-of-corpus problem. Using virtual examples, the annotated corpus can be extended in a fast, efficient and easy way. The experimental results show that the introduction of edit-distance produces some improvements in protein name recognition performance. And the model, which is trained with the corpus expanded by virtual examples, outperforms the model trained with the original corpus. According to the proposed methods, we finally achieve the performance 75.80 in F-measure (71.89 % in precision, 80.15 % in recall) in the experiment of protein name recognition on GENIA corpus (ver. 3.0).

Introduction

오늘날 생물학 분야의 활발한 연구 결과로 다양한 정보들이 매우 빠른 속도로 창출되고 있다. 그러나 이와 같은 정보들의 양이 매우 방대하기 때문에 효율적인 처리 및 이용이 점차 어려워지고 있는 실정이다. 이러한 정보들 중 특히 논문 등의 텍스트 정보를 보다 효율적으로 검색 및 처리하고 이용할 수 있도록 하기 위해 생물학 분야에 대한 텍스트 마이닝 기술에 대한 관심이 높아지고 있다.

기본적으로 텍스트 마이닝을 위해서는 다양

본 연구는 ETRI 위탁연구과제 지원으로 수행되었음.

한 언어처리 기술들이 필요하다. 이 중 단백질, 유전자 등의 명칭을 추출, 인식하는 개체명 인식 기술은 가장 기본적이고 핵심적인 중요 기술 중 하나이다 [1, 2]. 특히 단백질명은 전체 생물학적 개체명 개수의 약 80% 가량을 차지하므로 단백질명 인식 성능이 전체 개체명 인식 성능에 큰 영향을 미친다는 점에서 매우 중요하다. 또한 생물학적으로 중요한 정보인 단백질-단백질 상호작용의 추출에 있어서 높은 수준의 단백질명 인식 성능은 필수적인 요소라는 점에서도 매우 중요하다 [9, 10].

생물학 분야 텍스트 대상의 개체명 인식에 있어서 가장 큰 어려움 중 하나로 각 개체명들이 매우 다양한 철자 이형태(spelling variant form)를 가진다는 점을 들 수 있다. 개체명 사전이 잘 구축되어 있다 하더라도 모든 이형태를 포함시키는 것은 어려우므로, 이미 사전에 존재하는 개체명이라 하더라도 텍스트 상에서 다른 형태로 나타날 수 있으므로 인식에 어려움을 겪게 된다. 본 논문에서는 edit-distance를 SVM 기반 개체명 인식에 도입하여 이러한 문제를 해결하는 방법을 제안한다.

또한 학습 말뭉치의 규모는 기계학습 기반의 인식 성능에 크게 영향을 미치는 요소 중 하나이다. 개체명이 태깅된 학습 말뭉치의 양은 매우 제한적이므로 이를 효율적으로 증가시킬 수 있다면 인식 성능 향상에 큰 효과를 얻을 수 있다 [6, 7]. 본 논문에서는 가상 예제를 이용하여 학습 말뭉치의 양을 자동적인 방식에 의해 빠르고 효과적으로 증가시키는 방법을 제안한다.

본 논문은 다음과 같이 구성된다. 먼저 edit-distance 자질을 이용하는 SVM 기반의 개체명 인식 방법론에 대해 설명한다. 다음으로 가상 예제를 이용하여 학습 말뭉치를 확장하는 방법에 대해 설명한다. 이어서 제안된 방법론에 대한 단백질명 인식 실험 결과를 보이고, 마지막으로 본 연구에 대한 결론을 제시한다.

Related Works

철자 이형태 문제의 해결을 위해 Yi 등은 일반적인 HMM 기반 개체명 인식 모델에 edit-distance를 도입할 경우 기본 모델에 비해 F-measure 기준으로 약 3 % 가량의 성능 향상을 얻을 수 있음을 보였다 [6]. Yi 등은 edit-distance 값이 observation probability와 관련이 있음을 이용하여 HMM 기반 모델의 수식을 수정하여 edit-distance를 도입하였다. 그러나 SVM 기반 인식 모델에 edit-distance를 도입하고자 하는 경우 HMM 기반 인식 모델과는 달리 전체 모델의 수식에 대한 수정이 직관적으로 이루어지기가 매우 어렵다는 차이가 있다.

또한 Tsuruoka와 Tsujii는 edit-distance를 이용하여 개체명 후보를 추출한 뒤 추출된 후보를 필터링(filtering)하는 방법을 이용하여 단백질명 인식에 대해 F-measure 기준으로 70.2의 성능을 보였다 [10]. 이 때 edit-distance를 후보 선정에 이용하기 위해 입력 대상 전체에 대해 edit-distance를 계산하게 된다. 그러나 edit-distance의 가장 큰 단점이 계산 복잡도가 높다는 것임을 고려할 때, 이러한 방법은 계산량이 매우 많아 인식 속도에 문제가 생길 가능성이 있음을 알 수 있다. 이와는 달리 본 논문에서 제안하는 방법의 경우 SVM이 학습에 이용하는 데이터 양을 필터링을 통해 미리 제한시키는 등의 과정을 통해 실제 edit-distance가 계산

되는 경우는 전체 인식 대상의 극히 일부분에 불과하므로 속도 상으로 큰 손실을 입지 않고 성능 향상의 효과를 보는 것이 가능하다.

최소 edit-distance 자질을 사용한 SVM 기반의 개체명 인식

단백질명 외부 영역 단어의 필터링

학습에 이용되는 텍스트 내용에는 단백질명을 구성하는 단어보다 그렇지 않은 외부 영역 단어들이 상대적으로 훨씬 많이 포함되어 있다. SVM의 특성 상 이러한 불균형적인 학습 데이터의 분포는 분류 결과의 재현률을 떨어뜨리는 요인이 될 수 있다. 이러한 문제점을 해소하기 위하여 단위명사구 분리기(base noun phrase chunker)를 이용하여 명사구 영역에 포함되지 않는 단어는 단백질명을 구성하는 단어일 가능성이 낮다고 보고 학습 대상에서 제외시켰다. 이 과정을 통해 GENIA corpus [3] 3.01 버전의 총 490,455개 단어(token)들 중 약 40%를 걸러내고 총 280,266개 단어가 남았다.

이와 같이 필터링 과정으로 학습 데이터의 개수를 크게 줄인 결과, 학습 데이터 개수의 제곱에 비례하는 SVM의 학습 소요 시간을 크게 감소시키는 효과도 얻을 수 있었다. 더불어 학습 결과로 나오는 지지 벡터(support vector)의 수도 크게 감소하여 지지 벡터 개수에 비례하는 인식 시간도 상당량 단축시킬 수 있었다.

Edit-distance (ED)

문자열 X와 Y간의 edit-distance는 문자의 삽입, 삭제, 치환의 연산을 통하여 X를 Y로 변환할 때의 operation sequence의 가중치 총합으로 정의된다 [4]. Edit-distance는 두 문자열 간의 유사도 척도로 기능할 수 있기 때문에 염기 서열 및 아미노산의 유사도 측정, 문자인식, 철자 수정 등 다양한 분야에서 활용되어 왔으며, 특히 철자 수정 등의 분야에서는 철자 이형태의 처리에 이용되어 왔다 [5].

개체명 인식에 edit-distance를 도입할 경우, 개체명인지 판단해야 할 대상 X'가 사전의 개체명 X에 대해 충분히 작은 edit-distance 값을 가진다면 X'가 X의 한 이형태라고 판단하는 근거로 이용할 수 있으므로, 이형태 인식 문제 해결에 도움이 된다 [6].

SVM에서의 edit-distance 도입

본 논문에서는 단백질명의 분할 및 인식 문제를 대상 문장의 각 단어에 대해 적합한 분류 태그를 부여하는 classification 문제로 정의하고 인식모델을 만들었다. 이 때 분류 태그는 BIO 표현 양식을 따른다.

분류를 위한 기본 자질로는 단어 표층형(surface word), 단어의 형태 자질(orthographic feature of surface word), 품사 태그, 접두사, 접미사, 그리고 대상 단어 앞 단어의 분류 태그를 이용하였다.

Edit-distance를 SVM에 도입하기 위해 기본 자질에 edit-distance 자질을 추가하였다. 이 때 자질 값의 계산을 모든 가능한 입력 단어 열(word sequence)에 대해 수행할 경우 계산량이 지나치게 많아지며 실제 자질로서의 역할을 제대로 수행할 수 없다고 판단하였다. 실제 자질 값의 계산은 다음과 같은 과정에 의해서 필요한 경우에 대해서만 이루

어졌다. 먼저 임시 문자열을 빈 문자열로 초기화한 뒤, 입력의 각 단어에 대해 앞 단어의 개체명 분류 태그가 'B-PROTEIN'이거나 'I-PROTEIN'이면 현 위치 단어를 임시 문자열에 추가하고, 그 밖의 태그일 경우라면 현 단어를 임시 문자열에 덮어쓴다. 임시 문자열로부터 개체명 사전의 각 엔트리에 대한 edit-distance 중 최소값을 계산하고 이를 임시 문자열의 길이로 나누어 정규화(normalize)하여 해당 단어의 edit-distance 자질 값으로 이용하였다. Edit-distance를 계산할 때 [10]에서 제시된 단백질명 인식에 대한 cost function을 이용하여 단백질명에 대한 특징이 보다 많이 반영될 수 있도록 하였다.

가상 예제를 이용한 학습 말뭉치 확장

학습 데이터의 양을 증가시키기 위해 문자 인식이나 문서 분류 분야에서 이용되는 방법 중 하나로, 태깅된 데이터로부터 인공적으로 생성한 가상 예제를 이용하는 방법이 있다 [8].

기본적으로 단백질명은 명사이므로, 문장에 나타난 어떤 단백질명 X를 다른 단백질명 Y로 바꾼다 하더라도 문장의 기본 구조는 그대로 유지된다. 가상 예제는 이러한 점을 이용하여 생성하는데, 원래 말뭉치에서 단백질명을 만나면 해당 위치 단어 열을 사전에 있는 단백질명 중 임의로 하나를 선택하여 대치해 넣는다. 이러한 과정을 n번 반복하면 원래의 말뭉치가 약 (n+1)배로 확장되는 효과를 얻을 수 있다. 이렇게 자동으로 확장된 말뭉치는 원래 말뭉치보다 다양한 문맥 정보들을 포함하게 되므로 개체명 인식 성능 향상에 도움을 얻을 수 있다.

Experiment and Results

먼저 edit-distance 자질을 추가한 SVM 기반의 단백질명 인식 모델의 성능을 평가하기 위하여 GENIA corpus 버전 3.01의 총 2,000개 초록 중 무작위로 뽑은 1,600개를 학습용으로, 나머지 400개를 테스트용으로 실험을 수행하였다. 표 1에 제시된 결과를 보면 edit-distance 자질을 사용한 모델이 기본 자질만을 이용한 모델보다 정확도와 재현률 모두 성능 증가를 보여 f-measure 상으로 약 2 % 가량의 성능 향상을 보였음을 확인할 수 있다.

자질 유형	정확도 (%)	재현률 (%)	F-measure
기본	72.04	73.21	72.52
기본+ED	72.42	76.83	74.56

표 1 : Edit-distance의 효과.

가상 예제를 추가한 경우의 성능을 확인하기 위하여 앞서 사용한 학습 말뭉치의 각 문장당 4개의 가상 예제를 추가하여 만든 말뭉치로 인식 모델을 학습하고 앞서의 실험과 같은 테스트 데이터로 모델 성능을 평가하였다. (표 2). GENIA corpus만을 학습한 경우에 비하여 가상 예제를 추가하여 학습한 경우, 기본 자질만을 이용한 모델과 edit-distance 자질을 이용한 모델 모두 비록 정확도 면에서 약간의 손실을 입기는 하였으나 재현률이 상당히 높은 상승을 보여 F-measure 기준으로 성능 향상되었음을 확인할 수 있었다.

자질 유형	정확도 (%)	재현률 (%)	F-measure
기본	71.02	78.34	74.50
기본+ED	71.89	80.15	75.80

표 2 : 가상예제 추가 후 성능.

Conclusion

본 논문에서는 철자 이형태로 인한 생물 분야 개체명 인식의 어려움을 해소하기 위해 edit-distance를 도입하였다. 특히 분류에 있어서 우수한 성능을 보이는 것으로 알려진 SVM에 edit-distance를 자질로서 이용하는 방법을 제시하였다. 더불어 기계학습 기반의 인식 방법의 성능에 크게 영향을 미치는 학습 말뭉치 규모를 효율적으로 확장하기 위하여 가상 예제를 이용한 자동적 말뭉치 확장 방법을 제안하였다.

제안한 방법론에 대해 인식 성능 평가 실험을 수행한 결과, edit-distance 자질의 도입이 SVM 기반의 인식 모델의 성능을 향상시킴을 보였으며, 가상 예제를 통해 자동으로 확장된 말뭉치로 학습할 경우 추가적인 성능 향상을 보임을 통해 제안된 말뭉치 확장 방법의 효과를 입증하였다.

본 논문의 방법론은 단백질명 이외에 DNA, RNA 등의 분류에 속하는 개체명으로 비교적 손쉽게 확장이 가능하다. 현재 인식 대상에 DNA명을 추가하는 방법에 대해 기본적인 실험이 이루어진 상태이며, 그 외에 RNA, cell type, cell line 등에 대한 확장도 진행 중에 있다. 또한 GENIA corpus 이외에도 Yapex corpus 등으로 실험 대상을 확장하여 edit-distance 및 가상예제의 효과를 보다 확실히 입증하기 위한 과정을 거칠 예정이다.

References

[1] K. Fukuda, T.Tsunoda, A. Tamura and T.Takagi. Toward information extraction: Identifying protein names from biological papers. In Proceedings of PSB 1998, 1998.

[2] Ki-Joong Lee, Young-Sook Hwang and Hae-Chang Rim. Two-Phase Biomedical NE Recognition based on SVMs. In Proceedings of ACL 2003 Workshop on Natural Language Processing in Biomedicine, 2003.

[3] T. Ohta, Y. Tateisi, J. Kim, H. Mima and J. Tsujii. The genia corpus: An annotated research abstract corpus in molecular biology domain. In Proceedings of HLT 2002, 2002.

[4] R. A. Wagner and M. J. Fisher. The string-to-string correction problem. Journal of the Association for Computing Machinery, 21(1):168-173, Jan. 1974, 1974.

[5] F. J. Damerau. A technique for computer detection and correction of spelling errors. Communications of the ACM, 7(3):171-176, 1964.

[6] Eunji Yi, Gary Geunbae Lee, Soo-Jun Park. HMM-based protein name recognition with edit-distance using automatically annotated corpus. In Proceedings of the workshop on BioLINK text data mining SIG: Biology literature, information and knowledge, ISMB 2003, 2003.

[7] Juhui An, Seungwoo Lee, Gary Geunbae Lee. Automatic acquisition of Named Entity Tagged Corpus from World Wide Web. In Proceedings of ACL2003, July 2003.

[8] P. Niyogi, F. Girosi and T. Poggio. Incorporating prior information in machine

learning by creating virtual examples. In Proceedings of IEEE, volume 86, pages 2196-2207, 1998.

[9] K. Yamamoto, T. Kudo, A. Konagaya and Y. Matsumoto. Protein Name Tagging for Biomedical Annotation in Text. In Proceedings of the workshop on Natural Language Processing in Biomedicine, ACL 2003, 2003.

[10] Y. Tsuruoka and J. Tsujii. Boosting Precision and Recall of Dictionary-Based Protein Name Recognition. In Proceedings of the workshop on Natural Language Processing in Biomedicine, ACL 2003, 2003.