생의학 도메인에서 약어 중의성 해결을 위한 최적 자질의 규명

임호건 서희철 김선호 임해창
고려대학교 컴퓨터학과
{hglim, hcseo, shkim, rim}@nlp.korea.ac.kr

Identifying Optimum Features for Abbreviation Disambiguation in Biomedical Domain

Ho-Gun Lim Hee-Cheol Seo Seonho Kim Hae-Chang Rim
Dept. of Computer Science & Engineering, Korea University

요 약

생의학 도메인에서 약어 중의성 해결이란 생의학 문서에 나타난 약어의 원래 형태(long form)를 판별하는 작업이다. 본 논문은 생의학 도메인에서 약어 중의성 해결에 적합한 자질들을 실험적으로 탐색하는데 목적이 있다. 이를 위해서 약어 중의성 해결에 사용할 문맥을 전역 문맥(topical context)과 지역 문맥(local context)으로 구분하고, 각각의 문맥에서 스테밍(stemming), 불용어 제거, 품사 부착 등의 과정을 통해서 다양한 자질들을 고려하도록 한다. 생의학 도메인에서 약어 중의성 해결을 위한 실험 자료의 부족을 해결하기 위해서, 학습 자료와 평가 자료를 자동으로 구축했으며, 평가를 위한 약어로는 기존 연구에서 사용된 두 가지 약어 목록을 사용했다. 또한 단순 베이지언 모델(Naive Bayesian Model)을 이용해서 각 자질들의 유용성을 평가하였다. 실험 결과, 전역 문맥이 지역 문맥보다 더 좋은 성능을 보였으며, 전역 문맥에서는 불용어만을 제거한 경우가 각각의 평가 자료에서 94.2%와 96.2%로 가장 좋은 결과를 보였으며, 전역 문맥과 지역 문맥을 함께 사용하는 경우에 각각의 평가 자료에서 1.8%와 0.3%의 성능 향상이 있었다.

1. 서 론

최근 생의학 분야의 급속한 성장과 함께 대량의 관련 논문들이 양산되면서, 생의학 문서들에서 필요한 정보를 추출할 수 있는 방법에 관한 연구가 활발히 진행되고 있다. 이를 위해서 생의학 문서를 구성하는 다수의 전문 용어의 의미 분석이 절실히 요구되며, 더불어 전문 용어의 한 가지 형태인 약어 처리에 관한 관심이 고조되고 있다.

약어란 하나 이상의 단어열을 축약한 문자열로, 약어가 표현하는 하나 이상의 단어열을 약어의 원래 형태(long form)라 한다[1]. 생의학 분야에서는 다수의 전문 용어가 약어로 표현되며, 많은 수의 약어들이 두 가지 이상의 원래 형태를 가진다. 실제로 UMLS[1]에 등재된 약어의 33%가 중의성을 갖는다[2]. 또한, MEDLINE[2]에 출현하는 약어의 81.2%가 중의성을 가지며 평균 16.6개의 원래 형태를 갖는다[3].

약어 중의성 해결은 생의학 분야에서의 정보검색과 정보추출에서 필수적이다. 예를 들어, 정보검색 시스템에서 사용자가 'respiratory syncytial virus'를 의미하는 약어 'RSV'를 질의에 포함한 경우, 사용자의 질의 의도를 파악하기 위해서는 'RSV'의 원래 형태를 판별할 수 있어야 한다. 그리고 정보 추출에서는 전문 용어를 인식하고 분류하는 작업이 중요한데, 약어 중의성 해결만으로 전문 용어를 분류할 수 있다. 예를 들어, 약어 'APC'는 'antigen presenting cells'과 'activated protein c' 등의 원래 형태를 가지며, 각각은 세포와 단백질에 속한다. 그러므로 'APC'의 원래 형태를 알면 'APC'가 세포를 의미하는지, 단백질을 의미하는지를 판별할 수 있다.

1) UMLS(Unified Medical Language System)란, 생의학 분야에 쓰이는 다양한 어휘들을 집적해놓은 지식 데이터베이스를 말한다. www.nlm.nih.gov/research/umls

2) MEDLINE은 다양한 생의학 문서의 요약(abstract)만으로 구성된 대용량의 텍스트 데이터베이스이다.

그런데 이렇게 중의성이 높고 빈번하게 발생하는 약어의 원래 형태를 판별하는데 있어 문제는 하나의 문서에서 약어와 약어의 원래 형태를 동시에 기술하는 경우가 많지 않다는 데 있다. 따라서 약어의 중의성을 해결하는 문제는 간단한 규칙을 통해 해결할 수 있는 단순한 문제가 아니며, 대부분의 기존 연구는 기계 학습 기법을 통해서 약어 중의성 해결을 시도하고 있다. 그러나 약어 중의성 해결을 위한 기계 학습 기법에 적합한 자질들에 대해서는 아직까지 심도 있게 연구되지 않았다. 본 논문에서는 기계 학습 기법으로 생의학 분야에서 약어 중의성 해결을 할 때, 고려할 수 있는 자질들을 살펴보고, 약어 중의성 해결에 적합한 자질들을 판별하고자 한다.

2장에서는 약어 중의성 해결을 위한 관련 연구에 대해 알아보고, 3장에서는 약어 중의성 해결에 있어서 유용한 자질을 찾기 위한 접근 방법에 대해서 설명하고, 4장에서는 실험 결과를 보이며, 5장에서는 결론을 맺는다.

2. 관련 연구

본 장에서는 생의학 분야에서 수행했던 약어 중의성 해결에 관한 관련 연구를 살펴보고, 본 논문의 접근 방법과 유사한 단어 의미 중의성 해결에서의 관련 연구를 설명한다.

[Liu et al, 2001][4]은 약어 중의성 해결을 위해 주변 단어, 단어의 방향, 단어의 위치에 기반을 둔 세 가지 자질을 사용했으며, 문맥의 크기를 변경하면서 실험하였다. 기계 학습 기법으로는 단순 베이지언 모델(Naive Bayesian Model), 결정 리스트(decision lists), 표본 기반 학습(Exemplar-based learning)를 사용했다. 12개의 약어에 대해서 Clinical Data Repository[3]와 MEDLINE abstract로부터 자동으로 구축한 학습 자료와 평가 자료에서 실험을 했다.

[Liu et al, 2002][5]은 약어 중의성 해결을 위해 스테밍 처리한 단어들을 자질로 사용했다. 단순 베이지언 모델을 이용해서 중의성을 해결했으며, 학습자료와 평가 자료는 UMLS의 개념 관계(conceptual relation)와 "one sense per discourse[6]" 가정을 이용하여 MEDLINE으로부터 자동 구축하였으며, 11개 약어에 대해서 실험하였다.

[Pakomov 2002][7]는 약어와 함께 나타난 단어와 섹션 단위 문맥(section level context)[4]을 자질로 사용했다. 최대 엔트로피 모델(Maximum Entropy Model)을 적용하였으며, 학습 자료와 평가 자료는 Mayo Clinic에서 작성된 환자 관련 문서인 임상자료(clinical data)에서 자동으로 추출하였으며, 중의성이 높은 6개의 약어를 선별하여 약어 중의성 실험에 사용하였다. 추가적으로 약어와 공기하는 단어들과 약어의 원래 형태가 공기하는 단어가 유사함을 실험적으로 보임으로써 약어의 원래 형태를 가지고 학습 자료를 자동으로 만들 수 있음을 보였다. 본 논문에서도 이에 근거하여 실험에 사용하는 학습 자료와 평가 자료를 구축한다.

[Yu et al, 2003][8]은 약어 주변 네 개의 단어들을 자질로 사용했다. 기계 학습 기법으로는 지지벡터 기반 학습 알고리즘(support vector machine)을 사용하고, 중의성을 해결할 때 "one sense per discourse" 가정을 함께 이용했다. 실험을 위한 학습 자료와 평가 자료는 기존의 연구들과 동일하게 MEDLINE으로부터 자동으로 구축하였다. 또한 기존 연구들과의 비교를 위해서 [Pakomov, 2002]와 [Liu et al, 2002]에 사용된 약어 목록으로 실험을 했다. 본 연구에서도 동일한 약어 목록을 대상으로 실험하였다.

본 논문에서 살펴보는 약어 중의성 해결을 위한 자질 연구와 유사한 연구로 [Leacock et al., 1998][9]의 연구가 있다. [Leacock et al., 1998]은 단어 의미 중의성 해결에 적합한 자질들을 조사하기 위해서, 단어 의미 중의성 해결을 위한 문맥을 전역 문맥(topical context)과 지역 문맥(local context)으로 구분했다. 전역 문맥은 대상 단어가 나타난 문장과 앞/뒤 문장에 나타난 내용어(content word)로 구성되고, 지역 문맥은 대상 단어와 세 단어 이내에 위치한 단어, 품사, 위치 정보로 구성된다. 단순 베이지언 모델을 이용해서 전역 문맥과 지역 문맥의 유용성을 명사, 동사, 형용사 단어에 대해서 살펴보았으며, 실험 결과 명사는 전역 문맥에서, 동사와 형용사는 지역 문맥에서 가장 좋은 결과를 보였으며, 전역 문맥과 지역 문맥을 함께 사용하는 경우에 가장 좋은 결과를 보였다.

앞서 언급했듯이 생의학 분야에서 약어 중의성 해결을 위한 연구가 활발하게 진행되고 있지만, 대부분의 연구가 아직까지는 매우 단순한 자질만을 이용하고 있으며, 이용하는 자질에 대한 평가도 제대로 되고 있지 않다. 따라서 본 논문에서는 생의학 분야의 약어 중의

3) New York Presbyterian Hospital (NYPH)의 전자 진료 데이터를 말한다.

4) clinical data의 각 문서에는 문서의 범주가 부착되어 있는데, 이를 section level context라 정의함.

성 해결에 유용한 자질들을 분석하고자 한다. [Leacock et al., 1998]의 전역 문맥과 지역 문맥을 더 세분하여 모두 10가지 자질 유형을 구성하고, 각 유형별로 약어 중의성 해결에서 기여도를 살펴보았다. 나아가 개별 자질의 조합이 약어 중의성 해결에 미치는 영향에 대해서도 알아본다.

3. 접근 방법

본 절에서는 자질의 유용성을 평가하기 위해서 본 논문에서 접근한 방법에 대해서 설명한다. 먼저 자질의 유용성을 평가하기 위한 평가 자료와 학습 자료 구축에 대해서 기술하고, 약어 중의성 해결에서 사용되는 자질들을 설명한다. 마지막으로 약어 중의성 해결을 위해서 사용한 기계 학습 기법인 단순 베이지언 모델에 대해서 설명한다.

3.1. 학습 자료와 평가 자료의 자동 구축

기계 학습 기법을 이용해서 약어 중의성을 해결하기 위해서는 학습 자료가 필요하다. 그리고 자질의 유용성을 평가하기 위해서는 평가 자료가 필요하다. 그러나 생의학 분야에서 약어 중의성 해결을 위한 공개된 실험 자료가 없기 때문에 대부분의 기존 연구에서는 실험 자료를 자동으로 구축하는 방법을 제시하고 있다. 본 논문에서도 기존 연구와 같이 실험 자료 즉, 학습 자료와 평가 자료를 자동으로 구축한다.

문서에서 약어는 다양한 형태로 기술된다. 약어와 약어의 원래 형태가 '원래 형태(약어)' 형식으로 함께 기술되는 경우도 있고, 약어만 기술되는 경우도 있고, 약어 없이 원래 형태로만 기술되는 경우도 있다. 이와 같은 기술 형태에 근거해서 평가 자료와 학습 자료를 구축할 수 있다. 즉 '원래 형태(약어)'의 형태로 나타난 문장은 약어의 원래 형태를 정확하게 알 수 있으므로, 이런 문장들을 수집해서 실험 자료를 만들 수 있다. 그러나 RNA와 DNA와 같이 많이 알려진 약어는 원래 형태를 함께 기술하지 않는 경향이 있다. 그러므로 '원래 형태(약어)' 형식만으로는 학습 자료와 평가 자료 모두를 위한 충분한 자료를 수집할 수 없다. 본 논문에서는 '원래 형태(약어)' 형식은 평가 자료로 수집하고 학습 자료는 별도의 방법으로 수집하였다. 학습 자료는 기존의 연구 방법들과 유사한 방법으로 [Pakomov, 2002]의 "약어와 약어의 원래 형태는 유사한 단어들과 공기한다."는 가정에 근거하여 약어의 원래 형태만 나타난 문장으로 구성하였다. MEDLINE으로부터 평가 자료와 학습 자료를 자동으로 구축하는데 사용한 조건들은 다음과 같다.

〈1〉 평가 자료

1) 'LF (AB)' 형태가 나타난 문맥
2) 'AB'가 나타난 문맥 (MEDLINE abstract 내에서 'LF (AB)' 형태가 나타나고, LF는 AB의 유일한 LF이어야 함.).

〈2〉 학습 자료

1) 'LF'만 나타난 문맥

여기서 AB는 약어, LF는 약어의 원래 형태를 의미한다. 평가 자료를 구축하는 두 번째 조건에는 "one sense per discourse"와 유사한 "one long form per abstract"라는 가정이 적용된다. "***one long form per abstract***" 란, "하나의 초록(abstract)내에 나타나는 약어는 하나의 원래 형태를 갖는 경향이 있다"는 것을 뜻한다. 이 가정에 따라 특정 초록 내의 동일한 'AB'에 대해서 'AB'를 포함한 문맥들을 평가 자료에 추가한다.

자동으로 구축한 학습 자료와 평가 자료는 약어의 원래 형태가 나타난 문장과 앞, 뒤 문장을 포함하는 세 문장으로 추출된 인스턴스들로 구성된다. 실험 자료 추출을 위한 평가 목록은 기존 연구에서 사용했던 평가 목록과 동일한 약어 목록([Pakomov, 2002] [Liu et al, 2002])을 이용한다.

다음 [그림 1], [그림 2]는 MEDLINE abstract 중에서 약어 'RSV'가 포함된 문서들 중 하나이고, 위의 알고리즘을 통해 이들 문서에서 'RSV'의 학습 자료와 평가 자료 인스턴스가 각각 추출되는 예를 보인다.

> A radioimmunoassay was developed that can detect and quantitate 3 ng or more of the avian RNA tumor virus reverse transcriptase. The assay detected no antigenic sites in **Rous sarcoma virus** alpha virions or in virions of a murine RNA tumor virus. About 70 molecules of reverse transcriptase were found per virion of avian myleloblastosis virus with this assay or with an assay based on antibody inhibition of enzymatic activity. ...

[그림 1] 학습 자료 추출 예

[그림 1]에서 밑줄 친 부분은 약어 'RSV'의 원래 형태 중에서 'rous sarcoma virus'가 출현한 경우로 원래 형태만이 나타난 경우에는 알고리즘에 의해 원래 형태를 포함한 문장과 앞/뒤 문장의 세 문장이 다음과 같이

학습 자료로 추출된다.

〈문장1〉 *A radioimmunoassay was developed that can detect and quantitate 3 ng or more of the avian RNA tumor virus reverse transcriptase.*

〈문장2〉 *The assay detected no antigenic sites in Rous sarcoma virus alpha virions or in virions of a murine RNA tumor virus.*

〈문장3〉 *About 70 molecules of reverse transcriptase were found per virion of avian myleloblastosis virus with this assay or with an assay based on antibody inhibition of enzymatic activity.*

[그림 2]는 평가 자료를 추출하는 예를 보여준다. [그림 2]에서 밑줄 친 부분은 약어 'RSV'의 원래 형태 중에서 'respiratory syncytial virus'가 '원래 형태 (약어)' 형태로 나타난 경우이다. 또한 이후의 'RSV'는 "one long form per abstract" 가정에 의해 마찬가지로 평가 자료로 구성된다. [그림 2]에서는 다음과 같이 두 개의 평가 자료가 추출된다.

〈문장1〉 *Batches of commercial fetal bovine serum, described by the suppliers as antibody-free, all contained antibody to bovine syncytial virus (BSV) when tested by indirect immunofluorescence.*

〈문장2〉 *Antibody to bovine respiratory syncytial virus (RSV) was not detected in these sera.*

〈문장3〉 *Twenty-four percent of individual fetal bovine sera contained antibody to BSV, and 14% contained antibody to RSV when tested by indirect immunofluorescence.*

〈문장1〉 *Antibody to bovine respiratory syncytial virus (RSV) was not detected in these sera.*

〈문장2〉 *Twenty-four percent of individual fetal bovine sera contained antibody to BSV, and 14% contained antibody to RSV when tested by indirect immunofluorescence.*

> Batches of commercial fetal bovine serum, described by the suppliers as antibody-free, all contained antibody to bovine syncytial virus (BSV) when tested by indirect immunofluorescence. Antibody to bovine **respiratory syncytial virus (RSV)** was not detected in these sera. Twenty-four percent of individual fetal bovine sera contained antibody to BSV, and 14% contained antibody to **RSV** when tested by indirect immunofluorescence. ...

[그림 2] 평가 자료 추출 예

이와 같은 방법으로 [Pakomov, 2002]와 [Liu et al, 2002]에서 사용된 각각의 약어 목록에 대해서 구축한 학습 자료와 평가 자료는 [표 1], [표 2]와 같다.

3.2. 자질 집합

본 논문에서는 약어 중의성 해결을 위한 문맥을 전역 문맥과 지역 문맥으로 구분한다. 각 문맥은 품사 부착, 스테밍, 불용어 제거 등의 처리를 통해서 다음과 같이 10가지 자질로 세분된다.

〈1〉 전역 문맥(topical contexts)

- -문서 내에서 약어와 공기하는 큰 윈도우(문서 전체) 내의 문맥
- -구성
  1) TW : 약어와 공기하는 단어
  2) STW : 스테밍 처리한 TW
  3) SRTW : 불용어를 제거한 TW
  4) SSRTW : 스테밍 처리하고 불용어를 제거한 TW
  5) OCW : 약어와 공기하는 내용어(open-class word)
  6) SOCW : 스테밍 처리한 OCW

약어	원래 형태	평가자료 인스턴스 수	학습자료 인스턴스 수
BD	bundle	0	18176
	twice a day	0	4031
	band	0	69119
	계	0	91326
INF	infective	0	14644
	infant	0	68068
	inferior	5	55157
	interferon	161	49730
	infusion	9	197275
	infected	42	183833
	infection	18	485112
	계	235	1053819
NR	nonresponder	8	920
	nonreactive	14	1686
	no report	0	865
	no reponse	0	1
	nerve root	0	3489
	nurse	0	34346
	no refill	0	0
	no recurrence	0	4018
	no radiation	0	258
	normal range	12	13347
	계	34	58930
PA	periarteritis	0	418
	plasma aldosterone	688	4670
	pantothenic acid	119	488
	physician assistant	54	119
	pernicious anemia	115	727
	paranoia	3	395
	pyruvic acid	26	569
	panic attack	1	399
	paternal aunt	0	41
	procainamide	443	3360
	pulmonary artery	2235	29524
	pathology	3	55150
	polyarthritis	11	2338
	pseudomonas aeruginosa	429	14166
	polyarteritis	0	1691
	계	4127	114055
PN	periarteritis nodosa	23	336
	positional nystagmus	11	286
	parenteral nutrition	1189	9266
	polyneuritis	1	285
	pyelonephritis	71	6011
	polyneuropathy	62	5263
	peripheral nerve	236	10343
	peripheral neuropathy	218	4642
	polyarteritis nodosa	86	1288
	pneumonia	33	44821
	penicillin	7	23581
	계	1937	106122
RA	right atrial	734	9527
	rheumatic arthritis	31	43
	refactory anemia	3	1
	right atrium	531	6731
	right arm	0	1352
	radioactive	4	27689
	renal artery	133	12293
	rheumatoid arthritis	31600	19559
	계	33036	77195
총 계		39369	1501447

[표 1] 약어 목록 A([Pakomov, 2002])의 학습 자료와 평가 자료 크기

약어	원래 형태	평가자료 인스턴스 수	학습자료 인스턴스 수
ACE	antegrade colonic enema	13	14
	adrenocortial extract	0	0
	amsacrine cytarabine etoposide	0	0
	doxorubicin cyclophosphamide etoposide	0	0
	doxorubicin cyclophosphamide	0	0
	angiotensine converting enzyme	4	2
	acetylcholinesterase	68	12570
	계	85	12586
APC	activated protein c	0	1
	aphidicholin	2	7
	atrial premature complexes	8	23
	adenomatous polyposis coli	2186	367
	antigen presenting cells	764	906
	계	2960	1304
ASP	aspartate	175	18442
	aspartylglycine	0	0
	aspartic acid	18	3948
	asparaginase	44	878
	antisocial personality	74	1133
	ankylosing spondylitis	7	3377
	계	318	27778
BSA	bovine serum albumin	7961	5790
	body surface area	1237	3333
	계	9198	9123
CSF	cerebrospinal fluid	39820	20265
	colony stimulating factors	143	248
	cytostatic factor	156	57
	competence and sporulation factor	10	0
	계	40129	20570
EMG	exomphalos macroglossia gigantism	0	0
	electromyography	2759	3854
	electromyographs	12	41
	electromyogram	2582	1630
	계	5353	5525
IBD	inflammatory bowel disease	5868	5112
	irritable bowel syndrome	0	2501
	계	5868	7613
MAS	mccune albright syndrome	0	0
	meconium aspiration syndrome	320	262
	macandrew alcoholism scale	0	2
	계	320	264
RSV	respiratory syncytial virus	7743	1788
	rous sarcoma virus	1482	1462
	계	9225	3250
VCR	vincristine	2182	7223
	videocassette recorder	3	8
	vanadyl ribonucleoside complex	0	3
	계	2185	7234
총 계		75641	95247

[표 2] 약어 목록 B([Liu et al, 2002])의 학습 자료와 평가 자료 크기

〈2〉 지역 문맥(local contexts)

-약어 주변에 근접해 나타난 작은 윈도우(약어의 앞,뒤 세 단어) 내의 문맥.

-구성

1) LW : 약어와 공기하는 주변 단어
2) LWP : LW + 단어의 위치 정보
3) LWT : LW + 단어의 품사 정보
4) BW : LW의 바이그램(bigram)

스테밍은 Porter's stemmer[10]를 이용하며, 불용어를 제거하기 위해서 PubMed 불용어 목록[1]을 사용하며, 품사 부착을 위해서는 고려대학교 품사 부착기[2]를 이용한다.

[그림 2]에서 추출한 평가 자료에 품사 부착한 결과와 평가 자료에서 추출되는 10가지 자질들은 다음과 같다.

〈1〉 품사 부착된 평가 자료

〈문장1〉 Batches/NNS of/IN commercial/JJ fetal/JJ bovine/JJ serum/NN ,/, described/VBN by/IN the/DT suppliers/NNS as/IN antibody-free/NN ,/, all/DT contained/VBN antibody/NN to/TO bovine/JJ syncytial/JJ virus/NN (/(BSV/NN)/) when/WRB tested/VBN by/IN indirect/JJ immunofluorescence/NN ./.

〈문장2〉 Antibody/NN to/TO bovine/JJ respiratory syncytial virus (RSV)/RSV#1 was/VBD not/RB detected/VBN in/IN these/DT sera/NN ./.

〈문장3〉 Twenty-four/CD percent/NN of/IN individual/JJ fetal/JJ bovine/JJ sera/NNS contained/VBD antibody/NN to/TO BSV/NN ,/, and/CC 14/CD %/NN contained/VBD antibody/NN to/TO RSV/NN when/WRB tested/VBN by/IN indirect/JJ immunofluorescence/NN ./.

〈2〉 10가지 자질들

1) TW={'batches', 'of', 'commercial', 'fetal', 'bovine', 'serum', ',', ..., 'antibody', 'to', 'bovine', 'was', 'not', 'detected', 'in', 'these', 'sera', ..., 'when', 'tested', 'by', 'indirect', 'immunofluorescence', '.'}
2) STW={'batc', 'o', 'commerc', 'feta', 'bovi', 'seru', ..., 'antibod', 't', 'bovi', 'w', 'no', 'detec', 'i', 'thes', 'ser', ..., 'whe', 'tes', 'b', 'indirec', 'immunofluores'}
3) SRTW={'batches', 'commercial', 'fetal', 'bovine', 'serum', ',', ..., 'antibody', 'bovine', 'not', 'detected', 'sera', ..., 'tested', 'indirect', 'immunofluorescence', '.'}
4) SSRTW={batc', 'commerc', 'feta', 'bovi', 'seru', ..., 'antibod', 'bovi', 'no', 'detec', 'ser', ..., 'tes', 'indirec', 'immunofluores'}
5) OCW={'batches', 'commercial', 'fetal', 'bovine', 'serum', 'described', ..., 'antibody', 'bovine', 'was', 'not', 'detected', 'these', 'sera', ..., 'tested', 'indirect', 'immunofluorescence'}
6) SOCW={'batc', 'commerc', 'feta', 'bovi', 'seru', 'descri', ..., 'antibod', 'bovi', 'w', 'no', 'detec', 'thes', 'ser', ..., 'tes', 'indirec', 'immunofluores'}
7) LW={'antibody', 'to', 'bovine', 'was', 'not', 'detected'}
8) LWP={'antibody/-3', 'to/-2', 'bovine/-1', 'was/1', 'not/2', 'detected/3'}
9) LWT={'antibody/NN', 'to/TO', 'bovine/JJ', 'was/VBD', 'not/RB', 'detected/VBN'}
10) BW={'antibody_to', 'antibody_bovine', 'antibody_was', 'antibody_not', 'antibody_detected', 'bovine_to', 'to_was', 'not_to', 'detected_to', 'bovine_was', 'bovine_not', 'bovine_detected', 'not_was', 'detected_was', 'detected_not'}

3.3. 단순 베이지언 분류기 (Naive Bayes classifier)

본 논문에서는 약어 중의성 해결을 위해 단순 베이지언 분류기를 사용한다. 이 분류기는 다음 수식에 의해서 약어 중의성을 해결한다.

$$
\begin{aligned}
LF' &= \arg\max_{LF_k} p(LF_k \mid C) \\
&= \arg\max_{LF_k} \frac{p(C \mid LF_k)\,p(LF_k)}{p(C)} \\
&= \arg\max_{LF_k} p(C \mid LF_k)\,p(LF_k) \\
&= \arg\max_{LF_k} [\log p(C \mid LF_k) + \log p(LF_k)] \\
&= \arg\max_{LF_k} [\sum_j \log p(v_j \mid LF_k) + \log p(LF_k)]
\end{aligned}
$$

[수식 1] 베이지언 분류기

5) www.ncbi.nlm.nih.gov/entrez/
6) HMM 기반의 품사 부착기를 GENIA corpus에서 학습 시키고 사용함. 품사 집합은 Penn treebank set에 기인한다.

[표 3] 평가 자료 A에서 개별 자질을 이용한 결과

약어	MFLF[7]	전역 문맥						지역 문맥			
		TW	STW	SRTW	SSRTW	OCW	SOCW	LW	LWP	LWT	BW
INF	68.51	70.64	68.51	69.36	68.51	68.51	66.81	67.23	68.51	62.13	71.49
NR	41.18	79.41	79.41	79.41	79.41	79.41	79.41	55.88	47.06	41.18	55.88
PA	54.16	89.80	89.68	91.01	91.06	90.60	90.43	72.57	60.33	66.83	72.01
PN	61.38	85.85	86.47	86.78	87.15	85.85	87.04	52.81	35.62	50.28	52.04
RA	95.65	94.37	94.15	95.21	95.00	95.08	95.00	86.34	67.52	84.70	86.45
avg.	89.41	93.32	93.14	94.18	94.03	93.99	93.95	83.10	65.18	80.96	83.13

[표 4] 평가 자료 B에서 개별 자질을 이용한 결과

약어	MFLF	전역 문맥						지역 문맥			
		TW	STW	SRTW	SSRTW	OCW	SOCW	LW	LWP	LWT	BW
ACE	80.00	95.29	95.29	95.29	95.29	95.29	95.29	85.88	80.00	84.71	85.88
APC	73.85	96.79	96.93	96.89	97.33	96.76	97.40	80.64	70.37	80.51	76.39
ASP	55.03	94.65	94.03	94.97	94.34	94.65	94.03	76.10	71.07	75.47	72.01
BSA	86.55	97.35	97.12	97.99	97.73	98.15	97.89	91.49	86.21	91.49	89.00
CSF	99.23	99.59	99.49	99.61	99.50	99.60	99.46	99.14	98.59	99.11	98.64
EMG	51.54	72.88	72.84	72.89	72.86	73.01	72.97	66.75	59.69	66.00	64.38
IBD	100.00	91.58	91.04	91.77	91.16	91.17	90.68	88.48	80.28	88.26	85.89
MAS	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.69	99.06	99.69	99.06
RSV	83.93	94.32	94.74	94.96	95.17	94.72	94.80	87.69	80.35	87.50	85.49
VCR	99.86	99.86	99.86	99.91	99.86	99.91	99.86	99.86	99.91	99.73	99.82
avg.	91.33	96.04	95.96	96.23	96.12	96.17	96.05	92.88	89.49	92.76	91.49

여기서 LF'은 분류할 약어의 적절한 원래 형태이고 C는 주어진 문맥을 뜻하며, LF_k는 약어의 원래 형태 중 특정 원래 형태인 k번째 원래 형태를 가리킨다. v_j는 문맥 윈도우 내의 각 자질들을 가리키며 앞서 언급한 자질 집합으로 구성된다. 계산의 편의를 위해 로그를 취한다. 학습 자료의 자료 부족 문제는 additive smoothing을 적용해 완화시킨다.

4. 실험 결과

본 장에서는 앞서 만들었던 평가 자료에 대해서 10가지 자질들의 성능을 평가한다. 평가 척도는 다음 수식에 나타난 micro average precision(mi)을 이용한다.

$$mi = \frac{\text{평가자료에서 올바르게 분류한 instance 수}}{\text{평가자료에 있는 모든 instance 수}}$$

[수식 2] micro average precision(mi)

평가 자료 A, B에 대해 각 자질별 실험 결과는 [표 3], [표 4]와 같다. 각 평가 자료의 실험 결과에서 전역 문맥과 지역 문맥의 결과를 비교하면, 전역 문맥이 지역 문맥보다 항상 더 좋은 결과를 보여줌을 알 수 있다. 이는 약어 중의성 해결에서 약어와 좀 떨어져 있는 단어도 약어 중의성 해결에 중요한 영향을 준다는 것을 의미한다. 그리고 전역 문맥 중에서는 불용어를 제거한 경우가 두 가지 평가 자료에서 가장 좋은 결과를 보여주었다. 기존 연구들에서 많이 사용한 스테밍된 단어는 오히려 단어를 그대로 사용하는 경우보다 낮은 결과를 보였다. 그리고 품사 정보를 이용해서 내용어를 사용하는 경우 역시 스테밍 처리한 경우보다 높은 결과를 보여주었다. 이는 약어 중의성 해결에서는 불용어를 배제하고 의미 있는 내용어들을 사용한 경우에 높은 성능을 기대할 수 있음을 의미한다.

지역 문맥의 경우에는 전역 문맥에 비해서 더 낮은 결과를 보였지만 그 중에서 바이그램을 사용하는 경우(평가 자료 A)와 단어를 그대로 사용하는 경우(평가

7) 여기서 MFLF(most frequent long form)은 평가 자료에서 가장 빈도수가 높은 원래 형태로 분류한 성능이다.

자료 B)에 90% 정도의 성능을 보임을 알 수 있다. 즉, 지역 문맥만을 사용하더라도 뛰어난 성능을 보임을 알 수 있다. 그러나 지역 문맥에서 위치 정보를 함께 사용하는 경우에는 평가 자료 A에서 현저히 낮은 성능을 보였다. 품사 정보 역시 평가 자료 B에서 유용하게 사용되었지만 평가 자료 A에서 그렇지 못했다.

다음으로는 가장 좋은 성능을 보였던 전역 문맥 자질과 지역 문맥 자질을 조합하여 평가한다. 전역 문맥으로는 불용어 제거한 경우를 사용하며, 지역 문맥으로는 단어 바이그램과 단어 자체를 사용한 경우를 고려한다. 두 가지 평가 자료에서의 실험 결과는 [표 5], [표 6]과 같다. 전역 문맥과 지역 문맥을 조합한 경우가 각각의 문맥을 사용하는 경우보다 성능의 향상을 보였다. 이는 생의학 분야에서 약어 중의성 해결에서 전역 문맥과 지역 문맥을 함께 사용하는 것이 가장 좋은 자질임을 의미한다.

[표 5] 평가 자료 A에서 조합 자질을 이용한 결과

약어	MFLF	전역 문맥 + 지역 문맥	
		SRTW + BW	SRTW + LW
INF	68.51	80.00	72.77
NR	41.18	82.35	79.41
PA	54.16	91.28	91.69
PN	61.38	85.13	87.30
RA	95.65	97.28	96.34
avg.	89.41	95.94	95.25

[표 6] 평가 자료 B에서 조합 자질을 이용한 결과

약어	MFLF	전역 문맥 + 지역 문맥	
		SRTW + BW	SRTW + LW
ACE	80.00	95.29	95.29
APC	73.85	94.80	97.13
ASP	55.03	94.65	94.65
BSA	86.55	98.40	98.48
CSF	99.23	99.74	99.72
EMG	51.54	73.72	73.57
IBD	100.00	93.22	93.05
MAS	100.00	100.00	100.00
RSV	83.93	95.28	95.41
VCR	99.86	99.86	99.86
avg.	91.33	96.47	96.55

5. 결론 및 향후 연구

지금까지 생의학 도메인에서 빈번히 나타나는 약어의 중의성 해결에 유용한 자질에 관해 실험적으로 알아보았다. 실험 결과, 생의학 도메인의 약어 중의성 해결에서 유용한 자질은 전역 문맥이며 그 중에서도 불용어를 제거한 단어 자질임을 알 수 있었다. 그리고 전역 문맥과 지역 문맥을 함께 사용하는 경우에 가장 좋은 성능을 보임을 알 수 있었다. 다시 말해, 전역 문맥과 지역 문맥의 조합이 가장 유용한 자질이라는 결론을 내릴 수 있다.

향후 연구로는 좋은 결과를 보인 전역 문맥의 크기가 언제일 때 가장 좋은 지 실험을 통해 알아볼 계획이며, 실험 결과에서 지역 문맥과의 조합이 가장 좋은 분류 성능을 보였기에 전역 문맥의 크기에 따른 조합 자질의 성능에 대해서도 분석해 볼 계획이다. 더불어, 단순 베이지언 모델 외의 다양한 기계 학습 기법에서 약어 중의성 해결을 위한 자질들의 성능을 살펴보고, 다른 도메인으로의 적용 또한 고려해 볼 것이다.

6. 참고 문헌

[1] Chang J., Scheutze H. and Altman R. "Creating an online dictionary of abbreviations from MEDLINE". JAMIA 2002. Vol. 9, No. 6, Pages 612-620.

[2] Hongfang Liu, Yves A. Lussier, Carol Friedman "A study of Abbreviation in the UMLS", Proc. AMIA 2001; 393-97

[3] Hongfang Liu, Alan R. Aronson, Carol Friedman "A study of Abbreviation in MEDLINE Abstracts", AMIA 2002

[4] Hongfang Liu, Yves A. Lussier, Carol Friedman "Disambiguating ambiguous biomedical terms in biomedical narrative texts : Unsupervised methods" , JBI 2001 Aug; 34 (4): 249-61

[5] Hongfang Liu, Stephen B. Johnson, Carol Friedman "Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS", JAMIA 2002

[6] Yarowsky, D. "Unsuperivsed word sense disambiguation rivaling supervised methods" ACL 33th, 189-196

[7] Pakomov S. "Semi-supervised Maximum Entropy based approach to acronym", ACL 40th, July, 2002, 160-167

[8] Zhonghua YU, Yoshimasa TSURUOKA, Jun'ichi TSUJII "Automatic resolution of ambiguous abbreviation in biomedical texts using Support Vector Machines and one sense per discourse hypotheses", SIGIR 2003

[9] Claudia Leacock, Martin Chodorow, George A. Miller "Using corpus statistics and WordNet relations for sense identification" Computational Linguist 24, 1:147-65

[10] Porter, M. F. "An Algorithm for Suffix Stripping", M.F. Porter(1980),_Program_, Vol. 14,No. 3, Pp. 130-137. Program 1980, 14, 130-137.