Presentation of the "Acronym Test dataset"

For WSD evaluation in the biomedical domain

Contact: Dr Zhang Zhenling
Associate Professor
KETR Lab – University of Liaocheng – China
Condillac Research Group – University Savoie Mont-Blanc – France
zhenling_lotus@sina.com
http://ketrc.com/

Four test datasets are created and referred to as T100, T150, T200, and T300. They are created for WSD evaluation in the biomedical domain. The instances in all datasets are the abstracts downloaded from Pubmed (https://pubmed.ncbi.nlm.nih.gov/). Each of these abstracts contains an ambiguous acronym from the set of 18 acronyms originally developed in (Liu H et al., 2001)[1] and is widely used in previous WSD studies in the biomedical domain. Each acronym, used as an ambiguous target term in evaluation, consists of 3 letters, and it is associated with between 2 and 4 extended forms which are considered candidate senses. The extended forms for each acronym are obtained by querying the UMLS file MRCONSO.RRF[2] or the Lexicon.txt file of NLM[3]. Table 1 shows the long forms of each acronym, that is, the candidate senses of each ambiguous acronym.

Table 2 provides a resume of the acronym datasets used for the WSD evaluation. It can be seen, the number of senses per term in T100 and T200 is fixed. That is, the first two datasets T100, T150 both have fixed numbers of senses for each ambiguous word, 2 for T100 and 3 for T150 respectively.

Table 3 shows the specific candidate senses (CS) and the number of instances (NoI) for each acronym in each test dataset. As shown in Table 3, all the 18 acronyms are present in T100. The last two datasets T200 and T300 are mixed length datasets, and each acronym has between 2 and 4 extended forms. For an insufficient number of their extended forms, the acronyms "ANA", "BPD", "EMG", and "RSV" are not present in the data set T150 whose sense length is 3, and the acronyms "ANA", "FDP" are not present in the data set T200 whose sense length is between 2 to 4, and the acronyms "ANA", "DIP", "FDP", and "PVC" are not present in the data set T300 whose sense length is also between 2 to 4.

Table 1. Long forms of each acronym

Acronyms	Candidate senses (Long forms) of each acronym)

[1] Liu H, Lussier Y A, Friedman C. Disambiguating ambiguous biomedical terms in biomedical narrative text: an unsupervised method[J]. Journal of biomedical informatics, 2001, 34(4): 249-261.

[2] https://www.nlm.nih.gov/research/umls/index.html

[3] https://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lexicon/2020/release/LEX/LEXICON

ANA	S1	C0002463 American Nurses Association
	S2	C0003243 antinuclear antibody
APC	S1	C0003315 antigen presenting cell
	S2	C0483249 activated protein C
	S3	C1879736 argon plasma coagulation
BPD	S1	C0006287 bronchopulmonary dysplasia
	S2	C0006012 borderline personality disorder
BSA	S1	C0487992 body surface area
	S2	C0036774 bovine serum albumin
	S3	C0053170 benzenesulfonic acid
CAT	S1	C0040405 computed axial tomography
	S2	C1413138 Catalase
	S3	C0008169 chloramphenicol acetyltransferase
CML	S1	C002347 chronic myelocytic leukemia
	S2	C0301896 cell mediated lympholysis
	S3	E0400741+E0512845 complement mediated lysis
CMV	S1	C0419012 controlled mechanical ventilation
	S2	C0520499 Cucumber mosaic virus
	S3	C0319114 canine minute virus
DIP	S1	C0238378 desquamative interstitial pneumonia
	S2	C1511878 Diagnostic Imaging Program
	S3	C2266648 death-inducing-protein
EMG	S1	C0013839 Electromyography/ electromyograph/ electromyogram
	S2	C0004903 exomphalos-macroglossia-gigantism
FDP	S1	C1418197 fibrocyte-derived protein
	S2	C0163275 Fibrin degradation product
	S3	C1828476 flexor digitorum profundus
LAM	S1	C0751674 lymphangioleiomyomatosis

	S2	C0518959 left atrial myxoma
	S3	E0208573+ E0579913 linear associative memory
MAC	S1	C1167383 membrane attack complex
	S2	C0024432 Macrophage
	S3	C1416956 MARCKS Gene
MCP	S1	C0285488 Membrane Cofactor Protein
	S2	C0152392 Middle cerebellar peduncle
	S3	E0573715+ E0758850 microchannel plate
PCA	S1	C0030625 passive cutaneous anaphylaxis
	S2	C0078944 patient controlled analgesia
	S3	C0149576 Posterior Cerebral Artery
	S4	C0429865 Principal Components Analysis
PCP	S1	C1547431 primary care provider
	S2	C3896098 peptidyl carrier protein
	S3	C0030855 Pentachlorophenol
PEG	S1	C0176751 percutaneous endoscopic gastrostomy
	S2	C0032483 Polyethylene glycol product
	S3	C3814457 paternally expressed gene
PVC	S1	C0151636 Premature Ventricular Contraction
	S2	C0032624 polyvinyl chloride
	S3	C1166848 prevacuolar compartment
RSV	S1	C0035236 respiratory syncytial virus
	S2	C0086943 Rous sarcoma virus

Table 2. A resume of the acronym test datasets

Test datasets	T100	T150	T200	T300
Ambiguous terms	18	14	16	14
Instances	1800	1923	3105	4200
Min/Max senses per term	2	3	2/4	2/4
Average senses	2	3	2.92	2.85

Table 3. Number of instances for each acronym in the test datasets

Acronyms	T100		T150		T200		T300	
	CS	NoI	CS	NoI	CS	NoI	CS	NoI
ANA	S1, S2	100						
APC	S1, S2	100	S1, S2, S3	150	S1, S2, S3	200	S1, S2, S3	300
BPD	S1, S2	100			S1, S2	200	S1, S2	300
BSA	S1, S2	100	S1, S2, S3	150	S1, S2, S3	200	S1, S2, S3	300
CAT	S1, S2	100	S1, S2, S3	150	S1, S2, S3	200	S1, S2, S3	300
CML	S1, S2	100	S1, S2, S3	150	S1, S2, S3	200	S1, S2, S3	300
CMV	S1, S2	100	S1, S2, S3	107	S1, S2, S3	200	S1, S2, S3	300
DIP	S1, S2	100	S1, S2, S3	107	S1, S2, S3	105	S1, S2, S3	
EMG	S1, S2	100			S1, S2	200	S1, S2	300
FDP	S1, S2	100	S1, S2, S3	150				
LAM	S1, S2	100	S1, S2, S3	104	S1, S2, S3	200	S1, S2, S3	300
MAC	S1, S2	100	S1, S2, S3	105	S1, S2, S3	200	S1, S2, S3	300
MCP	S1, S2	100	S1, S2, S3	150	S1, S2, S3	200	S1, S2, S3	300
PCA	S1, S2	100	S1, S2, S3	150	S1,S2,S3,S4	200	S1,S2,S3,S4	300
PCP	S1, S2	100	S1, S2, S3	150	S1, S2, S3	200	S1, S2, S3	300
PEG	S1, S2	100	S1, S2, S3	150	S1, S2, S3	200	S1, S2, S3	300
PVC	S1, S2	100	S1, S2, S3	150	S1, S2, S3	200	S1, S2, S3	
RSV	S1, S2	100			S1, S2	200	S1, S2	300
Total instances		1800		1923		3105		4200