NTCIR-17 MedNLP-SC Social Media Adverse Drug Event Detection: Subtask Overview

Shoko Wakamiya[1], Lis Kanashiro Pereira[1], Lisa Raithel[234], Katherine Yeh[4], Peitao Han[1], Seiji Shimizu[1], Tomohiro Nishiyama[1], Gabriel Herman Bernardim Andrade[1], Noriki Nishida[5], Hiroki Teranishi[5], Narumi Tokunaga[5], Philippe Thomas[2], Roland Roller[2], Pierre Zweigenbaum[4], Yuji Matsumoto[5], Akiko Aizawa[6], Sebastian Möller[23], Cyril Grouin[4], Thomas Lavergne[4], Aurélie Névéol[4], Patrick Paroubek[4], Shuntaro Yada[1], Eiji Aramaki[1]

1 NAIST, Japan, 2 DFKI Berlin, Germany, 3 TU Berlin, Germany, 4 Université Paris-Saclay, CNRS, LISN, France, 5 RIKEN, Japan, 6 NII, Japan

Social Media Adverse Drug Event Detection (SM-ADE) Subtask

- To identify a set of symptoms caused by a drug (ADE; 副作用) from short messages written by social media users
- Social media corpus covering 17 drugs x 22 symptoms in four languages (Japanese 🇯🇵 , English 🇬🇧 , German 🇩🇪 , and French 🇫🇷)

ID	Japanese 🇯🇵	English 🇬🇧	German 🇩🇪	French 🇫🇷	#tweets
01	アザチオプリン	Azathioprine	Azathioprin	Azathioprine	600
02	アスピリン	Aspirin	Aspirin	Aspirine	500
03	アミオダロン	Amiodarone	Amiodaron	Amiodarone	500
04	インフリキシマブ	Infliximab	Infliximab	Infliximab	500
05	コルヒチン	Colchicine	Colchicin	Colchicine	500
06	シクロスポリン	Cyclosporin	Cyclosporin	Cyclosporine	500
07	シクロフォスファミド	Cyclophosphamide	Cyclophosphamid	Cyclophosphamide	500
08	シスプラチン	Cisplatin	Cisplatin	Cisplatine	1000
09	ステロイド剤	Steroids	Steroide	Steroids	500
10	タクロリムス	Tacrolimus	Tacrolimus	Tacrolimus	1000
11	ミノサイクリン	Minocycline	Minocyclin	Minocycline	500
12	メサラジン	Mesalazine	Mesalazin	Mesalazine	1000
13	メトトレキサート	Methotrexate	Methotrexat	Méthotrexate	500
14	メトホルミン	Metformin	Metformin	Metformine	500
15	抗結核薬	Anti-tuberculosis drugs	Anti-Tuberkulose-Mittel	Médicaments antituberculeux	400
16	抗生剤	Antibiotics	Antibiotika	Antibiotiques	500
17	造影剤	contrast media	Kontrastmittel	agents de contraste	500

ID	Japanese 🇯🇵	English 🇬🇧	German 🇩🇪	French 🇫🇷	UMLS CUI
01	悪心	nausea	Übelkeit	nausées	C0027497
02	下痢	diarrhea	Diarrhöe	diarrhée	C0011991
03	倦怠感	fatigue	Erschöpfung	fatigue	C0015672
04	嘔吐	vomiting	Erbrechen	vomissements	C0042963
05	食欲不振	loss of appetite	Anorexie	anorexie	C0003123
06	腹痛	abdominal pain	Unterleibsschmerzen	douleur abdominale	C0000737
07	頭痛	headache	Kopfschmerzen	maux de tête	C0018681
08	発熱	fever	Fieber	fièvre	C0015967
09	間質性肺疾患	interstitial lung disease	Interstitielle Lungenerkrankung	maladies pulmonaires interstitielles	C0206062
10	肝障害	liver damage	Leberschädigung	problèmes de foie	C0023895
11	浮動性めまい	Dizziness	Drehschwindel	sensation vertigineuse	C0012833
12	疼痛	pain	Schmerz	douleur	C0030193
13	脱毛症	alopecia	Alopezie	alopécie	C0002170
14	鎮痛剤喘息症候群	analgesic asthma syndrome	Analgetisches Asthma-Syndrom	syndrome d'asthme analgésique	(C0004096)
15	腎障害	renal impairment	Nierenerkrankung	insuffisance rénale	C0022658
16	過敏症	hypersensitivity	Hypersensibilität	hypersensibilité	C0020517
17	不眠症	insomnia	Insomnie	insomnie	C0917801
18	便秘	constipation	Constipation	constipation	C0009806
19	骨髄機能不全	bone marrow dysfunction	Knochenmarkerkrankung	trouble de la moelle osseuse	C0005956
20	出血性膀胱炎	hemorrhagic cystitis	Hämorrhagische Zystitis	cystite hémorragique	(C0010692)
21	発疹	rash	Ausschlag	éruption cutanée	C0015230
22	口内炎	stomatitis	Stomatitis	stomatite	C0149745

Social Media Corpus

- Generated 11,000 short messages in Japanese 🇯🇵 using a pre-trained language model (T5)
  Each tweet was manually checked and annotated with a positive (1) or negative (0) label for each ADE symptom
- Translated the corpus into English 🇬🇧 , German 🇩🇪 , and French 🇫🇷 by machine translation (DeepL) with manual check
  All language share the same ADE symptom label(s) and each subset consists of 9,957 messages; 80% training (7,964 messages) and 20% test (1,993 messages)

train_id	text 🇬🇧	C0027497: nausea	C0011991: diarrhea	C0015672: fatigue	C0042963: vomiting	C0003123: loss of appetite	C0018681: headache	C0015967: fever	C0206062: interstitial lung disease	C0023895: liver damage	C0012833: dizziness	C0030193: pain	C0002170: alopecia	C0004096: analgesic asthma syndrome	C0022658: renal impairment	C0020517: hypersensitivity	C0917801: insomnia	C0009806: constipation	C0005956: bone marrow dysfunction	C0000737: abdominal pain	C0010692: hemorrhagic cystitis	C0015230: rash	C0149745: stomatitis
7905	<user_name> In my case, about a year after I started taking mesalazine, I lost my appetite and had abdominal pain and diarrhea.	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
7908	I've been to the dermatologist, I think it's a side effect of the minocycline, but my eyes are itchy and I'm getting headaches... I hope I can finish the dose soon and get better. <url>	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7909	I had an asthma attack this morning. It's painful, like when your inhaler (steroids) wears off... I will see the doctor tomorrow!!!! <url>	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Evaluation Metrics

- **Full:** The performance over ADE labels (0 or 1)
  - Exact Match Accuracy
  - Per ADE Label: Precision, Recall, and $F1$ score for each label (0 and 1) across samples and classes
- **Individual:** The performance across symptoms
  - Per Symptom Class: Precision, Recall, and $F1$ score for each class
- **Binary:** How well models can detect examples containing ADEs independent of symptoms

Participant Systems and Performance

- 8 teams finally submitted their results (+Baseline)
  - 5 of them challenged all languages and all teams challenged the English track
  - Most teams ensembled models from multiple seeds
- **Overall:** Desirable improvements in medical applications 👍
  Compared to our baseline, F1 improved by around ~5%; some improved ~10+% Recall in positive class
- **BUT:** All results are within the range of around 0.8 in $F1$

Approaches

Model	Extra data
SapBERT📋	no
XLM-RoBERTa	no
GPT-3.5, XLM-RoBERTa	data aug.
BERT, GPT-3.5, GPT-4	data aug.
BERT, XLM-RoBERTa	no
VADER, BERT	no
BERT, ClinicalDistilBERT📋	no
mBERT, RoBERTa, DeBERTa, XLM-RoBERTa	no
BERT, RoBERTa, XLM-RoBERTa	no

Results of the Binary Score for teams in each language track

Team	Metrics	Japanese 🇯🇵 ADE	Japanese non-ADE	English 🇬🇧 ADE	English non-ADE	German 🇩🇪 ADE	German non-ADE	French 🇫🇷 ADE	French non-ADE
AILABUD	Precision	0.57	**0.99**	0.58	**0.98**	0.56	**0.98**	0.55	**0.98**
	Recall	**0.98**	0.71	**0.97**	0.71	**0.96**	0.69	**0.97**	0.69
	F1	0.72	0.82	0.72	0.83	0.70	0.81	0.70	0.81
FRAG	Precision	0.78	0.94	0.76	0.93	0.76	0.92	0.76	0.92
	Recall	0.86	0.90	0.83	0.90	0.81	0.90	0.81	0.89
	F1	0.82	0.92	0.80	0.91	0.78	0.91	0.78	0.91
HPIDHC	Precision	0.79	0.94	0.77	0.94	**0.78**	0.92	0.78	0.92
	Recall	0.86	0.91	0.86	0.90	0.82	**0.91**	0.81	0.91
	F1	0.82	0.92	0.81	0.92	0.80	**0.92**	0.80	0.92
IMNTPU	Precision	–	–	0.74	0.91	–	–	–	–
	Recall	–	–	0.78	0.89	–	–	–	–
	F1	–	–	0.76	0.90	–	–	–	–
SRCB	Precision	**0.81**	0.94	**0.79**	0.94	0.75	0.97	**0.79**	0.93
	Recall	0.86	**0.92**	0.86	**0.91**	0.93	0.87	0.84	0.91
	F1	**0.83**	0.83	**0.82**	0.92	**0.83**	0.92	**0.82**	0.92
STIS	Precision	–	–	0.75	0.91	–	–	–	–
	Recall	–	–	0.78	0.90	–	–	–	–
	F1	–	–	0.77	0.90	–	–	–	–
TMUNLP	Precision	–	–	0.73	0.94	–	–	–	–
	Recall	–	–	0.86	0.87	–	–	–	–
	F1	–	–	0.79	0.90	–	–	–	–
VLP	Precision	0.77	0.93	0.76	0.92	0.75	0.91	0.76	0.92
	Recall	0.83	0.90	0.82	0.90	0.78	0.90	0.81	0.90
	F1	0.80	0.92	0.79	0.91	0.77	0.90	0.78	0.91
Baseline$_{XLM-R_{ALL}}$	Precision	0.76	0.92	0.75	0.92	0.73	0.90	0.73	0.91
	Recall	0.82	0.90	0.82	0.89	0.74	0.89	0.79	0.88
	F1	0.79	0.91	0.78	0.91	0.74	0.89	0.76	0.90

Results of the Per Symptom Class setting evaluation for all teams in the English track 🇬🇧

Team	Metrics	C0027497	C0011991	C0015672	C0042963	C0003123	C0018681	C0015967	C0206062	C0023895	C0012833	C0030193	C0002170	C0004096	C0022658	C0020517	C0917801	C0009806	C0005956	C0000737	C0010692	C0015230	C0149745
AILABUD	Precision	0.70	0.52	0.71	0.85	0.65	0.70	0.51	0.33	0.12	0.50	0.36	0.50	0.53	0.36	0.47	0.66	0.68	0.33	0.28	0.50	0.74	0.50
	Recall	0.99	0.95	0.98	1.00	0.92	1.00	0.96	1.00	1.00	1.00	1.00	0.83	0.88	0.94	0.80	0.93	0.85	0.97	1.00	0.98	1.00	0.85	0.95
	F1	0.82	0.68	0.82	0.92	0.76	0.83	0.67	0.50	0.21	0.67	0.51	0.64	0.68	0.50	0.63	0.74	0.80	0.50	0.44	0.67	0.79	0.66
FRAG	Precision	0.79	0.74	0.76	0.86	0.75	0.82	0.70	0.00	0.00	0.59	0.64	0.70	0.89	0.60	**0.77**	0.73	0.76	0.00	0.75	0.00	0.76	0.65
	Recall	0.92	0.82	0.91	0.82	0.88	0.88	0.66	0.00	0.00	0.77	0.54	0.88	0.94	0.60	0.82	0.71	0.94	0.00	0.86	0.00	0.76	0.77
	F1	0.85	0.78	0.83	0.84	0.81	0.85	0.68	0.00	0.00	0.67	0.59	0.78	0.92	0.60	0.79	0.72	0.84	0.00	0.80	0.00	0.76	0.71
HPIDHC	Precision	0.79	0.78	**0.78**	0.89	0.81	0.82	0.70	0.00	0.00	0.52	**0.68**	0.80	0.85	0.75	0.74	0.79	0.75	0.33	0.73	0.50	0.76	0.68
	Recall	0.94	0.88	0.89	0.77	0.90	0.95	0.74	0.00	0.00	0.85	0.62	**1.00**	0.94	0.60	0.71	0.56	0.87	0.50	0.90	0.50	0.79	0.68
	F1	0.86	0.82	0.83	0.83	0.85	0.88	**0.72**	0.00	0.00	0.65	**0.65**	**0.89**	0.89	0.67	0.73	0.66	0.81	0.40	0.81	0.50	0.78	0.68
IMNTPU	Precision	0.77	0.75	0.76	0.83	0.75	0.77	0.63	0.00	0.00	0.53	0.53	0.70	0.81	0.83	0.76	0.84	0.74	0.33	0.72	0.67	0.75	0.67
	Recall	0.91	0.83	0.89	0.86	0.87	0.95	0.58	0.00	0.00	0.69	0.44	0.88	0.94	**1.00**	0.68	0.47	0.84	0.50	0.75	0.50	0.64	0.64
	F1	0.83	0.79	0.82	0.84	0.80	0.85	0.61	0.00	0.00	0.60	0.48	0.78	**0.91**	0.72	0.60	0.79	0.40	0.73	0.57	0.69	0.65
SRCB	Precision	0.81	**0.80**	0.76	0.86	**0.82**	0.86	0.65	1.00	**0.50**	0.69	0.66	0.77	0.89	**1.00**	0.73	0.79	**0.85**	0.50	0.74	**0.80**	**0.81**	0.82
	Recall	0.96	0.93	0.93	0.86	0.90	0.95	0.74	0.50	0.50	0.85	0.53	0.88	0.89	0.80	0.86	0.56	0.94	0.50	0.91	1.00	0.76	0.82
	F1	0.88	**0.86**	0.84	0.86	**0.86**	**0.90**	0.69	0.67	**0.50**	**0.76**	0.58	0.82	0.89	0.89	0.79	0.66	**0.89**	0.50	0.82	**0.89**	0.78	0.82
STIS	Precision	0.81	0.78	0.73	0.77	0.81	**0.87**	0.52	0.00	0.00	0.65	0.65	**1.00**	0.00	0.74	0.80	0.81	0.00	0.73	0.00	0.74	**0.90**
	Recall	0.92	0.76	0.84	0.91	0.85	0.91	0.60	0.00	0.00	0.85	0.46	0.50	0.89	0.00	0.41	0.24	0.68	0.00	0.82	0.00	0.61	0.41
	F1	0.86	0.77	0.78	0.83	0.83	0.89	0.56	0.00	0.00	0.73	0.54	0.67	0.94	0.00	0.67	0.36	0.74	0.00	0.67	0.00	0.67	0.56
TMUNLP	Precision	0.76	0.75	0.74	0.79	0.74	0.79	0.64	1.00	0.00	0.55	0.58	0.70	0.94	0.80	**0.85**	0.76	0.50	0.62	0.60	0.70	0.64
	Recall	0.91	0.88	0.82	0.86	0.92	0.95	0.68	0.50	0.00	0.85	0.61	0.88	0.94	0.80	0.79	0.50	**1.00**	0.50	0.88	0.75	0.73
	F1	0.83	0.81	0.78	0.83	0.82	0.86	0.66	0.67	0.00	0.67	0.59	0.78	0.94	0.80	0.72	0.63	0.86	0.50	0.73	0.67	0.74	0.68
VLP	Precision	**0.84**	0.78	0.77	0.81	0.73	0.80	0.70	0.00	0.33	0.52	0.61	0.70	0.94	0.75	0.73	0.72	0.72	**0.67**	0.73	0.40	0.74	0.76
	Recall	0.93	0.88	0.91	0.77	0.90	0.91	0.62	0.00	0.50	0.85	0.51	0.88	0.94	0.60	0.79	0.62	0.84	1.00	0.82	0.50	0.61	0.73
	F1	0.88	0.83	0.84	0.79	0.81	0.86	0.66	0.00	0.40	0.65	0.56	0.78	0.94	0.67	0.76	0.67	0.78	**0.80**	0.77	0.44	0.67	0.74
Baseline$_{XLM-R_{ALL}}$	Precision	0.79	0.73	0.75	**0.90**	0.77	0.79	0.66	0.00	0.00	0.56	0.54	0.70	0.94	0.67	0.71	0.69	0.74	0.00	**0.76**	0.00	0.74	0.64
	Recall	0.93	0.83	0.82	0.86	0.85	0.91	0.55	0.00	0.00	0.77	0.60	0.88	0.94	0.40	0.79	0.53	0.94	0.00	0.82	0.00	0.61	0.73
	F1	0.85	0.78	0.79	0.88	0.81	0.86	0.60	0.00	0.00	0.65	0.57	0.78	0.94	0.50	0.75	0.60	0.83	0.00	0.79	0.00	0.67	0.68

Remaining Issues

- Generated messages are not perfect: how to do it better?
  Some are medically dubious and translated often "sound" Japanese, despite being English/German/French
- How to get the privacy free data?

If you have any suggestions, please contact us!
Web: https://sociocom.naist.jp/mednlp-sc/
E-mail: mednlp-sc@is.naist.jp