Real-MedNLP:

Overview of REAL document-based MEDical Natural Language Processing Task

Shuntaro Yada
Yuta Nakamura
Shoko Wakamiya
Eiji Aramaki

NAIST
東大病院 The University of Tokyo Hospital

Medical NLP Today

- Medical AI ≒ Medical **Image** AI

- Why NLP-based AI is not popular
  - Medical text data is alway small
    - Privacy Information
    - Language barrier

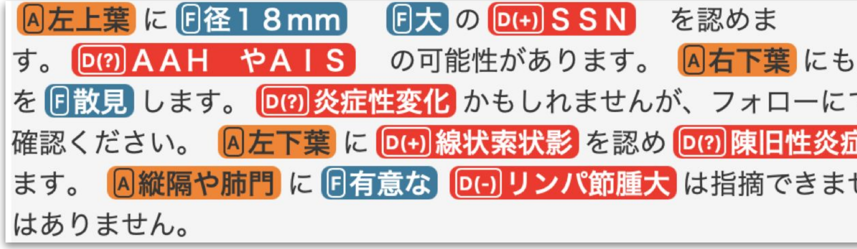
- Especially, **non-English medical NLP** is rare

Characterics of Our Task

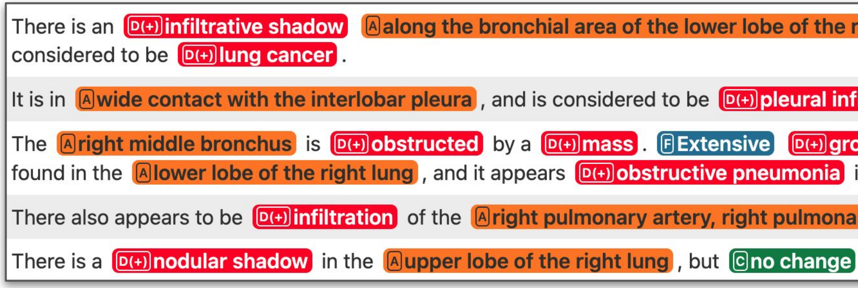
1. To provide High quality data-set
   - Real data (not dummy)
   - Closslingual (not English only)

2. To scope practical
   - Not only basic technology
     - Namerd Entity Recognition
   - ready-to-use applications
     - ADE detection
     - Case Identification

Japanese (JA)



English (EN)



Task & Language

- Two corpora x Three tasks x Two languages

		Care Report corpusMedTxt-CR Track		Radiology Report corpus MedTxt-RR Track	
Few-resources Named Entity Recognition; NER	Subtask 1 : Just 100 Training	Ja	En	Ja	En
	Subtask 2 : Guideline Learning	Ja	En	Ja	En
Application	Subtask 3 : Application (ADE* or CI**)	ADE-Ja	ADE-En	CI-Ja	CI-En

* Adverse Drug Event detection, ** Case Identification

Statistics of participants

- Although 19 teams registered, 9 teams submitted the results
- Balanced participation of international industry and academia

10 teams dropout

Number of registered teams: 19	Overseas: 13* (China, USA, Switzerland, Belgium, Germany)	Domestic (Japan): 7*
	Industry: 10	Academia: 9
Number of completed teams: 9	Overseas: 4* (China, USA, Switzerland)	Domestic (Japan): 6*
	Industry: 6	Academia: 3

*Since one team is composed of two countries, it is double-counted

Number of systems developed by each team
(85 systems by 9 teams)

	A	C	D	E	F	G	H	I	J	Total
Subtask1-CR-JA	2			1	4	1			4	12
Subtask1-CR-EN		2			4		5		4	15
Subtask1-RR-JA	2			1		1			4	8
Subtask1-RR-EN							3		4	7
Subtask2-CR-JA	1			1						2
Subtask2-CR-EN										0
Subtask2-RR-JA	1			1						2
Subtask2-RR-EN								1		1
Subtask3-CR-JA (ADE)				1	2					3
Subtask3-CR-EN (ADE)		10			2		6	1		19
Subtask3-RR-JA (CI)			1	1	1				1	4
Subtask3-RR-EN (CI)		10			1			1		12

Subtask 1 & 2

Task Definition = Named Entity Recognition

腸脛靭帯摩擦症候群を疑った変形性膝関節症患者 - 膝外側部痛に対するプレーティングアプローチによる介入の一症例 -

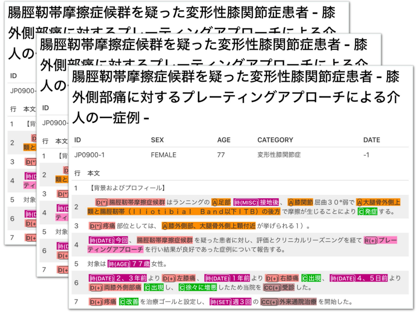
ID	SEX	AGE	CATEGORY	DATE
JP0900-1	FEMALE	77	変形性膝関節症	-1

行	本文
1	【背景およびプロフィール】
2	D(*) 腸脛靭帯摩擦症候群 はランニングの A 足部 時(MISC) 接地後、A 膝関節 屈曲３０°弱で A 大腿骨外側上顆と腸脛靭帯（Ｉｌｉｏｔｉｂｉａｌ　Ｂａｎｄ以下ＩＴＢ）の後方 で摩擦が生じることにより C 発症 する。
3	D(*) 疼痛 部位としては、A 膝外側部、大腿骨外側上顆付近 が挙げられる１）。
4	時(DATE) 今回、腸脛靭帯摩擦症候群 を疑った患者に対し、評価とクリニカルリーズニングを経て R(+) プレーティングアプローチ を行い結果が良好であった症例について報告する。
5	対象は 時(AGE) ７７歳 女性。
6	時(DATE) ２、３年前 より D(+) 左膝痛、時(DATE) １年前 より D(+) 右膝痛 C 出現、時(DATE) ４、５日前 より D(+) 両膝外側部痛 C 出現 し、C 徐々に増悪 したため当院を CC(+) 受診 した。
7	D(+) 疼痛 C 改善 を治療ゴールと設定し、時(SET) 週３回 の CC(+) 外来通院治療 を開始した。

- Diseases and symptoms <d>
- Anatomical entities <a>
- Features and measurements <f>
- Change <c>
- Time <timex3>
- Test <t-test/key/val>
- Medicine <m-key/val>
- Remedy <r>
- Clinical Context <cc>

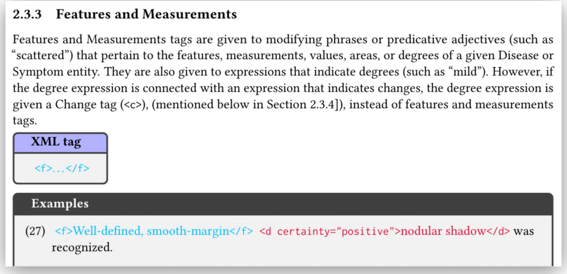
CR	<a>, <d>, <t-test/key/val>, <m-key/val>, <timex3>
RR	<a>, <d>, <t-test>, <timex3>

Subtask 1 – Just 100 Training



- Provide only 100-200 documents for training
- Standard few/low-resource NER setting

Subtask 2 – Guideline Learning


2.3.3 Features and Measurements

Features and Measurements tags are given to modifying phrases or predicative adjectives (such as “scattered”) that pertain to the features, measurements, values, areas, or degrees of a given Disease or Symptom entity. They are also given to expressions that indicate degrees (such as “mild”). However, if the degree expression is connected with an expression that indicates changes, the degree expression is given a Change tag (<c>), (mentioned below in Section 2.3.4]), instead of features and measurements tags.

XML tag
<f>...</f>

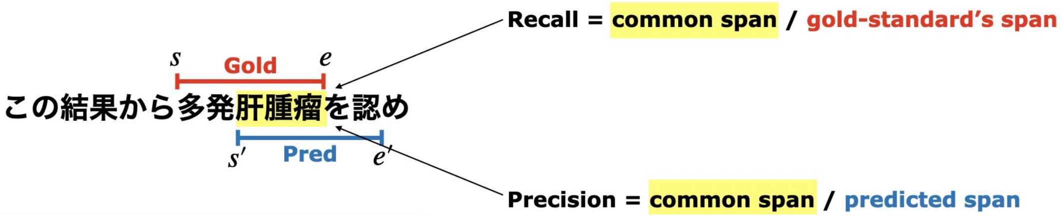
Examples

(27) <f>Well-defined, smooth-margin</f> <d certainty="positive">nodular shadow</d> was recognized.

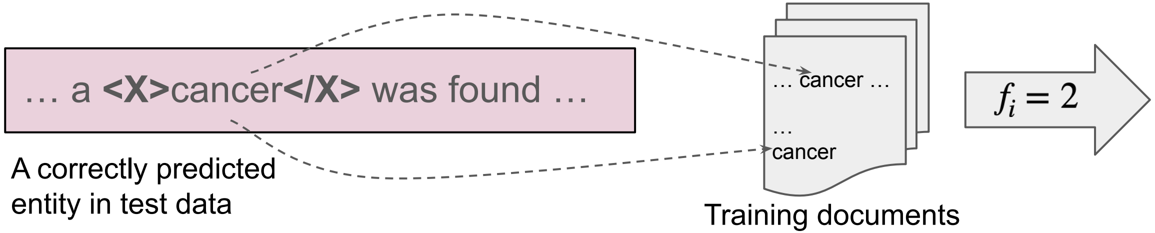
- Provide only the *guideline text* for human annotators
  - 30-40 example sentences annotated
- Can we teach a model as if it is a human?

Evaluation metrics

- Joint factor
  - span <X>cancer</X>
  - +label <d>cancer</d>
  - +label+mod <d mod=”positive”>cancer</d>
- Matching policy
  - exact
  - partial



- Frequency factor
  - Not weighted
  - Weighted – decrease the score according as the entity appeared once or more in training data



$$w_i = \frac{1}{\log_e(f_i + 1) + 1}$$

Weighting function

Overview (overall)

RR	**>>**	**CR**	RR ~ 0.9 vs. CR ~ 0.7 Radiology reports are written simpler than case reports
JA	>	EN	Japanese results are slightly better, but not a big difference
Partial	**>>**	**Exact**	At least ~10 points better in the partial match → “Important” parts of documents were still learnable
Normal	>	Weighted	Frequency weighting always decrease the scores
Subtask 1	>>	Subtask 2	Fewer training examples impacted

Overview (Subtask1-CR: Just 100 Case Reports)

- In JA, surprisingly, “just a plain BERT” (E1) worked best
- Simple data augmentation may rather decrease the performance

CR-JA

System ID	Exact match					
	span		+label		+label+mod	
	normal	weighted	normal	weighted	normal	weighted
A1	0.6388	0.5433	0.6133	0.5195	-	-
A2	0.6378	0.5425	0.6124	0.5188	-	-
E1	**0.6988**	**0.5995**	**0.6525**	**0.5550**	**0.5921**	**0.4993**
F1	0.6095	0.5112	0.5696	0.4737	0.5249	0.4333
F2	0.6497	0.5445	0.6076	0.5048	0.5602	0.4621
F3	0.5897	0.4987	0.5550	0.4650	0.5171	0.4315
F4	0.6179	0.5218	0.5813	0.4863	0.5420	0.4515
G1	0.6766	0.5754	0.6189	0.5198	-	-
J1	0.3361	0.2627	0.3088	0.2383	0.2591	0.1963
J2	0.3676	0.3057	0.3585	0.2968	0.3013	0.2459
J3	0.2745	0.2279	0.2656	0.2195	0.2247	0.1836
J4	0.2841	0.2399	0.2773	0.2334	0.2308	0.1910

CR-EN

System ID	Exact match					
	span		+label		+label+mod	
	normal	weighted	normal	weighted	normal	weighted
C1	0.4601	0.4117	0.4321	0.3850	-	-
C2	0.4697	0.4198	0.4371	0.3890	-	-
F1	0.5104	0.4501	0.4683	0.4092	0.4245	0.3701
F2	0.5292	0.4667	0.4860	0.4247	0.4406	0.3843
F3	0.5240	0.4634	0.4918	0.4326	0.4480	0.3938
F4	0.5473	0.4839	0.5145	0.4525	0.4696	0.4127
H1	0.6246	0.5513	0.5980	0.5255	0.5484	0.4809
H2	**0.6540**	**0.5813**	**0.6337**	**0.5616**	**0.5853**	**0.5181**
H3	0.6438	0.5719	0.6231	0.5515	0.5749	0.5080
H4	0.6190	0.5516	0.5933	0.5265	0.5452	0.4831
H5	0.6299	0.5620	0.6033	0.5364	0.5540	0.4917
J1	0.4882	0.4274	0.4556	0.3965	0.2957	0.2589
J2	0.5551	0.4925	0.5197	0.4589	0.3335	0.2950
J3	0.5503	0.4846	0.5116	0.4478	0.3263	0.2867
J4	0.5270	0.4652	0.4918	0.4317	0.3077	0.2705

Overview (Subtask1-RR: Just 100 Radiology Reports)

- Frequency weighting yielded larger performance drops than CR
    - Dataset contains template phrases more
- Domain-specific BERTs worked better as expected

RR-JA

System ID	Exact match					
	span		+label		+label+mod	
	normal	weighted	normal	weighted	normal	weighted
A1	0.1528	0.1185	0.1505	0.1165	-	-
A2	**0.9019**	**0.5264**	**0.8926**	**0.5181**	-	-
E1	0.8704	0.5052	0.8488	0.4871	**0.8079**	**0.4674**
G1	0.8932	0.5207	0.8703	0.4992	-	-
J1	0.5862	0.3232	0.5811	0.3191	0.4259	0.2550
J2	0.6055	0.3306	0.6022	0.3278	0.4363	0.2572
J3	0.5805	0.3151	0.5779	0.3127	0.4224	0.2480
J4	0.5715	0.3120	0.5674	0.3096	0.4216	0.2477

RR-EN

System ID	Exact match					
	span		+label		+label+mod	
	normal	weighted	normal	weighted	normal	weighted
H1	0.8296	0.5532	0.8260	0.5496	**0.7919**	**0.5262**
H2	**0.8302**	**0.5536**	**0.8266**	**0.5500**	0.7874	0.5231
H3	0.8140	0.5430	0.8061	0.5358	0.7719	0.5105
J1	0.7696	0.5049	0.7592	0.4957	0.6350	0.4107
J2	0.8068	0.5360	0.7997	0.5299	0.6707	0.4400
J3	0.7962	0.5265	0.7877	0.5192	0.6532	0.4264
J4	0.8000	0.5332	0.7895	0.5245	0.6545	0.4309

Subtask 2 – Guideline Learning

- Exact match resulted in an expected low score
- Partial match showed promising results

CR-JA

System ID	Exact match						Partial match					
	span		+label		+label+mod		span		+label		+label+mod	
	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted
A1	**0.4212**	**0.4146**	**0.3710**	**0.3644**	**0.3710**	**0.3644**	**0.7458**	**0.7379**	**0.6163**	**0.6091**	**0.6163**	**0.6091**
E1	0.3366	0.3326	0.2512	0.2474	0.1949	0.1912	0.6797	0.6738	0.4589	0.4547	0.3464	0.3424

RR-JA

System ID	Exact match						Partial match					
	span		+label		+label+mod		span		+label		+label+mod	
	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted
A1	**0.6638**	**0.6370**	**0.6485**	**0.6217**	**0.5133**	**0.4958**	**0.9106**	**0.8834**	**0.8843**	**0.8571**	**0.6864**	**0.6685**
E1	0.6557	0.6315	0.6255	0.6013	0.4668	0.4462	0.8961	0.8711	0.8289	0.8039	0.6094	0.5880

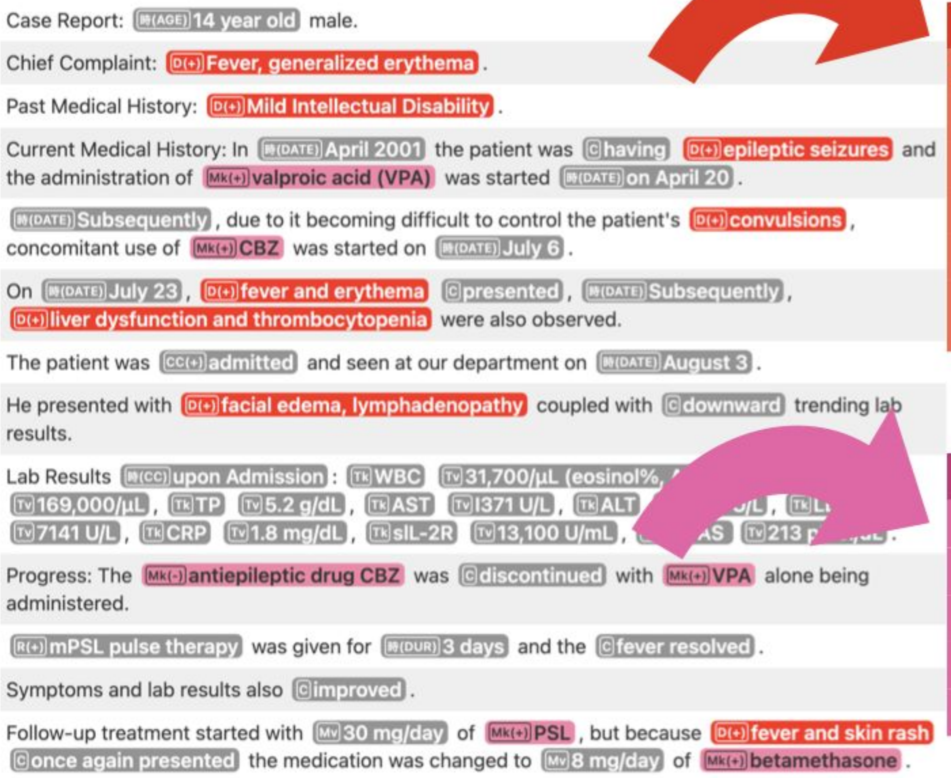
RR-EN

System ID	Exact match						Partial match					
	span		+label		+label+mod		span		+label		+label+mod	
	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted	normal	weighted
I1	0.5628	0.5546	0.5496	0.5422	0.5037	0.4968	0.8843	0.8726	0.8289	0.8179	0.7599	0.7495

Subtask 3 ADE

Table Slot Filling

For disease/medicine entities, predict the likelihood of being/triggering an ADE independently*


Case Report: 時(AGE) 14 year old male.
Chief Complaint: D(+) Fever, generalized erythema.
Past Medical History: D(+) Mild Intellectual Disability.
Current Medical History: In 時(DATE) April 2001 the patient was C having D(+) epileptic seizures and the administration of Mk(+) valproic acid (VPA) was started 時(DATE) on April 20.
時(DATE) Subsequently, due to it becoming difficult to control the patient's D(+) convulsions, concomitant use of Mk(+) CBZ was started on 時(DATE) July 6.
On 時(DATE) July 23, D(+) fever and erythema C presented, 時(DATE) Subsequently, D(+) liver dysfunction and thrombocytopenia were also observed.
The patient was CC(+) admitted and seen at our department on 時(DATE) August 3.
He presented with D(+) facial edema, lymphadenopathy coupled with C downward trending lab results.
Lab Results 時(CC) upon Admission: Tk WBC Tv 31,700/μL (eosinol%, Tv 169,000/μL, Tk TP Tv 5.2 g/dL, Tk AST Tv 1371 U/L, Tk ALT Tv 7141 U/L, Tk CRP Tv 1.8 mg/dL, Tk sIL-2R Tv 13,100 U/mL, Tv 213
Progress: The Mk(-) antiepileptic drug CBZ was C discontinued with Mk(+) VPA alone being administered.
R(+) mPSL pulse therapy was given for 時(DUR) 3 days and the C fever resolved.
Symptoms and lab results also C improved.
Follow-up treatment started with Mv 30 mg/day of Mk(+) PSL, but because D(+) fever and skin rash C once again presented the medication was changed to Mv 8 mg/day of Mk(+) betamethasone.

Disease	ADEval
Fever, generalized erythema	3
Mild Intellectual Disability	0
epileptic seizures	0
convulsions	0
fever and erythema	3
liver dysfunction and thrombocytopenia	3

= How likely is this disease (symptoms) an ADE?

3 – Definitely
2 – Probably
1 – Unlikely
0 – Unrelated

Medicine	ADEval
valproic acid (VPA)	1
CBZ	3
VPA	1
PSL	2
betamethasone	0

= How likely did this medicine trigger an ADE?

* do not consider ADE-causal relations

Results

- No ADEval=2 in test
- Better entity-level systems may not perform better in the report level
- How to capture local/global context seems important to solve this task
    - Classification?
    - NER?

CR-JA

System ID	ADEval=0			ADEval=1			ADEval=3			Report-level		
	P	R	F	P	R	F	P	R	F	P	R	F
E1	95.21	76.04	84.55	0.00	0.00	0.00	6.98	**52.94**	12.33	12.73	**77.78**	21.88
F1	95.76	**97.67**	**96.71**	0.00	0.00	0.00	12.50	11.76	12.12	**37.50**	66.67	**48.00**
F2	**96.05**	97.00	96.52	0.00	0.00	0.00	**27.59**	47.06	**34.78**	25.00	44.44	32.00

CR-EN

System ID	ADEval=0			ADEval=1			ADEval=3			Report-level		
	P	R	F	P	R	F	P	R	F	P	R	F
C1	95.70	94.94	95.32	20.00	5.26	8.33	62.50	26.32	37.04	22.22	66.67	33.33
C2	95.79	97.00	96.39	14.29	5.26	7.69	43.75	36.84	40.00	29.41	55.56	38.46
C3	95.95	93.52	94.72	6.25	5.26	5.71	28.57	21.05	24.24	19.35	66.67	30.00
C4	96.05	92.10	94.03	25.00	5.26	8.70	22.22	42.11	29.09	18.92	77.78	30.43
C5	95.87	95.26	95.56	0.00	0.00	0.00	56.25	47.37	51.43	25.93	77.78	38.89
C6	96.14	94.47	95.30	25.00	10.53	14.81	50.00	21.05	29.63	21.21	77.78	33.33
C7	95.67	94.31	94.99	0.00	0.00	0.00	33.33	26.32	29.41	19.35	66.67	30.00
C8	96.42	97.79	97.10	20.00	5.26	8.33	47.62	52.63	50.00	50.00	77.78	60.87
C9	96.35	91.79	94.01	0.00	0.00	0.00	23.81	52.63	32.79	18.92	77.78	30.43
C10	95.87	95.26	95.56	7.14	5.26	6.06	26.92	36.84	31.11	23.08	66.67	34.29
F1	96.53	96.68	96.61	0.00	**96.68**	0.00	31.25	52.63	39.22	25.00	55.56	34.48
F2	95.39	98.10	96.73	0.00	0.00	0.00	40.00	42.11	41.03	40.00	44.44	42.11
H1	96.57	97.95	97.25	14.29	5.26	7.69	60.00	63.16	61.54	50.00	66.67	57.14
H2	96.57	97.95	97.25	0.00	0.00	0.00	59.09	**68.42**	63.41	50.00	66.67	57.14
H3	96.28	98.10	97.18	0.00	0.00	0.00	60.00	63.16	61.54	50.00	55.56	52.63
H4	96.41	97.63	97.02	0.00	0.00	0.00	57.14	63.16	60.00	50.00	66.67	57.14
H5	95.88	**99.37**	**97.60**	0.00	0.00	0.00	78.57	57.89	**66.67**	**60.00**	33.33	42.86
H6	95.99	98.26	97.11	**33.33**	5.26	9.09	55.56	52.63	54.05	50.00	44.44	47.06
I1	**97.02**	97.63	97.32	30.00	31.58	**30.77**	**100.00**	26.32	41.67	50.00	**88.89**	**64.00**

Subtask 3 CI

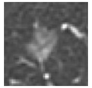
CI (Case Identification) = Clustering task

- Motivation: to recognize clinically similar documents without being confused by **textual similarity**
- Potential application: case retrieval, image-to-text evaluation

Clinically similar

Clinically different

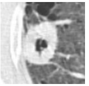
An 18mm ground-glass opacity is depicted in the left S1+2.



Ground-glass / sub-solid

A sub-solid nodule with a diameter of 18mm is seen in the upper left lobe.

A ***cavitated*** nodule with a diameter of 18mm is seen in the upper ***right*** lobe.



Cavitated

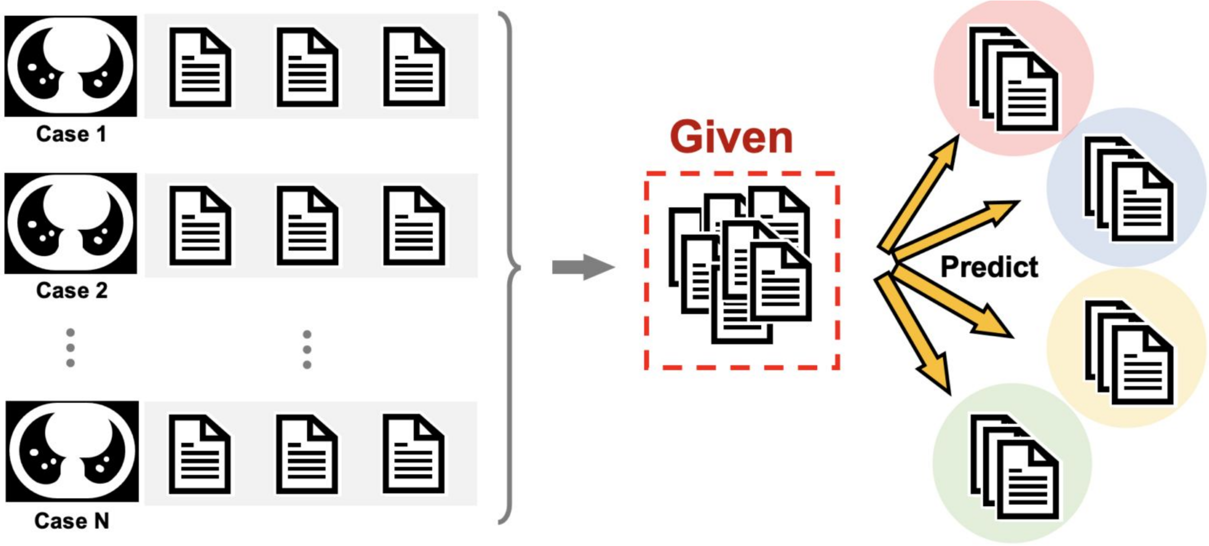
https://radiopaedia.org/cases/minimally-invasive-adenocarcinoma-of-the-lung-1?lang=us
https://radiopaedia.org/cases/cavitating-bronchogenic-carcinoma?lang=us

CI (Case Identification) = Clustering task

Data: radiology reports by nine radiologists

Goal: put the same case into the same cluster

15 cases
(training:test
= 8:7)



Evaluation metric: Normalized Mutual Information (NMI)

Results (RR-JA)

- Surprisingly, simple “similar or not similar” classification with BERT works best

System ID	NMI score	Method	
D1	0.3569	Bag-of-entity vectors	NER-based
E1	**0.5415**	**Binary document-pair classification with BERT**	Document representation
F1	0.1744	mBERT encoding + dimensionality reduction + K-means clustering	Document representation
J1	0.4161	Sentence classifications	Document representation
J1*	0.4622	Sentence classifications	Document representation

Results (RR-EN)

- System C1 achieved the best score with a pipeline method with rule-based approach & K-means clustering

System ID	NMI score	Method
C1	**0.8721**	**Heuristic + K-means clustering with SentenceBERT**
F1	0.2172	mBERT encoding + dimensionality reduction + K-means clustering
I1	0.7879	Named entity representations with BERT

Rule-based + Document representation

Document representation

NER-based

Results (Summary)

- NER-based: JA << EN
  - Maybe due to absence of well-organized Japanese medical ontology
- Document representation only << Pipeline approach
  - Suggesting importance of macroscopic & microscopic features

System ID	NMI score	Method
C1	**0.8721**	**Heuristic + K-means clustering with SentenceBERT**
F1	0.2172	mBERT encoding + dimensionality reduction + K-means clustering
I1	0.7879	Named entity representations with BERT
D1 (RR-JA)	0.3569	Bag-of-entity vectors

Rule-based + Document representation

Document representation

NER-based

NER-based

Conclusions

Conclusion

- RQ: Can we develop MedNLP applications with low resources?
  - YES (partly)
- NER
  - Promising performance for radiology reports (less diverse than case reports) even when only annotation guidelines are provided
- ADE
  - Fair, but discrepancy remains between entity- and document-level performance
- CI
  - High performance for English corpus: token- to document-level features may be needed

Conclusion

We look forward to hearing your presentations!

The comments on the next task is always welcomed

来年のタスクについてのご意見も歓迎です

Acknowledgement

Supported by **AIP-PRISM**, Japan **KEEPHA** project of JST, and **Y’s READING** inc.