Task 2: ShARe/CLEF eHealth Evaluation Lab 2014

Danielle L. Mowery[1], Sumithra Velupillai[2], Brett R. South[3], Lee Christensen[3], David Martinez[4], Liadh Kelly[5], Lorraine Goeuriot[5], Noemie Elhadad[6], Sameer Pradhan[7], Guergana Savova[7], and Wendy W. Chapman[3] ⋆

[1] University of Pittsburgh, PA, USA, dlm31@pitt.edu
[2] Stockholm University, Sweden, sumithra@dsv.su.se
[3] University of Utah, UT, USA, brett.south@hsc.utah.edu, leenlp@q.com,
wendy.chapman@utah.edu
[4] University of Melbourne and MedWhat (CA,USA), VIC, Australia,
davidm@csse.unimelb.edu.au
[5] Dublin City University, Ireland, Firstname.Lastname@computing.dcu.ie
[6] Columbia University, NY, USA, noemie.elhadad@columbia.edu
[7] Harvard University, MA, USA, sameer.pradhan@childrens.harvard.edu,
guergana.savova@childrens.harvard.edu

Abstract. This paper reports on Task 2 of the 2014 ShARe/CLEF eHealth evaluation lab which extended Task 1 of the 2013 ShARe/CLEF eHealth evaluation lab by focusing on template filling of disorder attributes. The task was comprised of two subtasks: attribute normalization (task 2a) and cue identification (task 2b). We instructed participants to develop a system which either kept or updated a default attribute value for each task. Participant systems were evaluated against a blind reference standard of 133 discharge summaries using Accuracy (task 2a) and F-score (task 2b). In total, ten teams participated in task 2a, and three teams in task 2b. For task 2a and 2b, the HITACHI team systems (run 2) had the highest performances, with an overall average average accuracy of 0.868 and F1-score (strict) of 0.676, respectively.

Keywords: Natural Language Processing, Template Filling, Information Extraction, Clinical Text

1 Introduction

In recent years, healthcare initiatives such as the United States *Meaningful Use* [1] and European Union *Directive 2011/24/EU* [2] have created policies and legislation to promote patient involvement and understanding of their personal health information. These policies and legislation have encouraged health care

⋆ DLM, SV, WWC led the task, WWC, SV, DLM, NE, SP, and GS defined the task, SV, DLM, BRS, LC, and DM processed and distributed the dataset, and SV, DLM, and DM led result evaluations organizations to provide patients open access to their medical records and advocate for more patient-friendly technologies. Patient-friendly technologies that could help patients understand their personal health information, e.g., clinical reports, include providing links for unfamiliar terms to patient-friendly websites and generating patient summaries that use consumer-friendly terms and simplified syntactic constructions. These summaries could also limit the semantic content to the most salient events such as active disorder mentions and their related discharge instructions. Natural Language Processing (NLP) can help by filtering non-active disorder mentions using their semantic *attributes* e.g., negated symptoms (*negation*) or uncertain diagnoses (*certainty*) [3] and by identifying the discharge instructions using text segmentation [4, 5].

In previous years, several NLP shared tasks have addressed related semantic information extraction tasks such as automatically identifying concepts - problems, treatments, and tests - and their related attributes (2010 i2B2/VA Challenge [6]) as well as identifying temporal relationships between these clinical events (2012 i2B2/VA Challenge [7]). The release of these semantically-annotated datasets to the NLP community is important for promoting the development and evaluation of automated NLP tools. Such tools can identify, extract, filter and generate information from clinical reports that assist patients and their families in understanding the patient's health status and their continued care. The ShARe/CLEF eHealth 2014 shared task [8] focused on facilitating understanding of information in narrative clinical reports, such as discharge summaries, by visualizing and interactively searching previous eHealth data (Task 1) [9], identifying and normalizing disorder attributes (Task 2), and retrieving documents from the health and medicine websites for addressing questions mono- and multi-lingual patients may have about the disease/disorders in the clinical notes (Task 3) [10]. In this paper, we discuss Task 2: disorder template filling.

2 Methods

We describe the ShARe annotation schema, the dataset, and the evaluation methods used for the ShARe/CLEF eHealth Evaluation Lab Task 2.

2.1 ShARe Annotation Schema

As part of the ongoing Shared Annotated Resources (ShARe) project [11], disorder annotations consisting of disorder mention span offsets, their SNOMED CT codes, and their contextual attributes were generated for community distribution. For 2013 ShARe/CLEF eHealth Challenge Task 1[12] the disorder mention span offsets and SNOMED CT codes were released. For 2014 ShARe/CLEF eHealth Challenge Task 2, we released the disorder templates with 10 attributes that represent a disorder's contextual description in a report including *Negation Indicator*, *Subject Class*, *Uncertainty Indicator*, *Course Class*, *Severity Class*, *Conditional Class*, *Generic Class*, *Body Location*, *DocTime Class*, and *Temporal*

Expression. Each attribute contained two types of annotation values: normalization and cue detection value. For instance, if a disorder is negated e.g., "*denies* nausea", the **Negation Indicator** attribute would represent nausea with a normalization value: *yes* indicating the presence of a negation cue and cue value: *start span-end span* for *denies*. All attributes contained a slot for a cue value with the exception of the *DocTime Class*. Each note was annotated by two professional coders trained for this task, followed by an open adjudication step.

From the ShARe guidelines[13], each disorder mention contained an **attribute** *cue* as a text span representing a non-default normalization value (*default normalization value)[8]:

Negation Indicator (NI): def. indicates a disorder was negated: *no, *yes*
Ex. "*No* cough."

Subject Class (SC): def. indicates who experienced a disorder: *patient, *family_member*, donor_family_member, donor_other, null, other
Ex. "*Dad* had MI."

Uncertainty Indicator (UI): def. indicates a measure of doubt about the disorder: *no, *yes*
Ex. "*Possible* pneumonia."

Course Class (CC): def. indicates progress or decline of a disorder: *unmarked, changed, increased, decreased, improved, worsened, *resolved*
Ex. "Bleeding *abated*."

Severity Class (SV): def. indicates how severe a disorder is: *unmarked, slight, moderate, *severe*
Ex. "Infection is *severe*."

Conditional Class (CO): def. indicates existence of disorder under certain circumstances: *false, *true*
Ex. "Return *if* nausea occurs."

Generic Class (GC): def. indicates a generic mention of disorder: *false, *true*
Ex. "Vertigo *while* walking."

Body Location (BL): def. represents an anatomical location: *NULL, *CUI: C0015450*, CUI-less
Ex. "*Facial* lesions."

DocTime Class (DT): def. indicates temporal relation between a disorder and document authoring time: *before*, after, overlap, before-overlap, *unknown

Ex. “Stroke in *1999.*”

Temporal Expression (TE): def. represents any TIMEX (TimeML) temporal expression related to the disorder: *none, *date*, time, duration, set
Ex. “Flu on *March 10.*”

2.2 Dataset

At the time of the challenge, the ShARe dataset consisted of 433 de-identified clinical reports sampled from over 30,000 ICU patients stored in the MIMIC (Multiparameter Intelligent Monitoring in Intensive Care) II database [14]. The initial development set contained 300 documents of 4 clinical report types - discharge summaries, radiology, electrocardiograms, and echocardiograms. The unseen test set contained 133 documents of only discharge summaries. Participants were required to participate in Task 2a and had the option to participate in Task 2b.

For Task 2a and 2b, the dataset contained templates in a “|” delimited format with: a) the disorder CUI assigned to the template as well as the character boundary of the named entity, and b) the default values for each of the 10 attributes of the disorder. Each template contained the following format [12]:

DD_DocName|DD_Spans|DD_CUI|Norm_NI|Cue_NI|
Norm_SC|Cue_SC|Norm_UI|Cue_UI|Norm_CC|Cue_CC|
Norm_SV|Cue_SV|Norm_CO|Cue_CO|Norm_GC|Cue_GC|
Norm_BL|Cue_BL|Norm_DT|Norm_TE|Cue_TE

For example, the following sentence, “The patient has an extensive thyroid history.”, was represented to participants with the following disorder template with default normalization and cue values:

09388-093839-DISCHARGE_SUMMARY.txt|30-36|C0040128|*no|*NULL|
patient|*NULL|*no|*NULL|*false|*NULL|
unmarked|*NULL|*false|*NULL|*false|*NULL|
NULL|*NULL|*Unknown|*None|*NULL

For Task 2a: Normalization, participants were asked to either keep or update the normalization values for each attribute. For the example sentence, the Task 2a **changes**:

09388-093839-DISCHARGE_SUMMARY.txt|30-36|C0040128|*no|*NULL|
patient|*NULL|*no|*NULL|*false|*NULL|
unmarked|*NULL|**severe**|*NULL|*false|*NULL|
C0040132|*NULL|**Before**|*None|*NULL

For Task 2b: Cue detection, participants were asked to either keep or update the cue values for each attribute. For the example sentence, the Task 2b **changes**:

09388-093839-DISCHARGE_SUMMARY.txt|30-36|C0040128|*no|*NULL|
patient|*NULL|*no|*NULL|*false|*NULL|
unmarked|*NULL|severe|**20-28**|*false|*NULL|
C0040132|**30-36**|Before|*None|*NULL

In this example, the Subject Class cue span is not annotated in ShARe since *patient is an attribute default.

2.3 Participant Recruitment and Registration

We recruited participants using listservs such as AMIA NLP Working Group, AISWorld, BioNLP, TREC, CLEF, Corpora, NTCIR, and Health Informatics World. Although the ShARe dataset is de-identified, it contains sensitive, patient information. After registration for task 2 through the CLEF Evaluation Lab, each participant completed the following data access procedure, which included (1) a CITI [15] or NIH [16] Training certificate in Human Subjects Research, (2) registration on the Physionet.org site [17], (3) signing a Data Use Agreement to access the MIMIC II data.

2.4 Evaluation Metrics

For Tasks 2a and 2b, we determined system performance by comparing participating system outputs against reference standard annotations. We evaluated overall system performance and performance for each attribute type e.g., *Negation Indicator*.

Task 2a: Normalization Since we defined all possible normalized values for each attribute, we calculated system performance using Accuracy as *Accuracy* = *count of correct normalized values* divided by *total count of disorder templates.*

Task 2b: Cue Detection Since the number of strings not annotated as attribute cues (i.e., ***true negatives (TN)***) is very large, we followed [18] in calculating F1-score as a surrogate for kappa. F1-score is the harmonic mean of recall and precision, calculated from true positive, false positive, and false negative annotations, which were calculated as follows:

true positive (TP) = the annotation cue span from the participating system overlapped with the annotation cue span from the reference standard
false positive (FP) = an annotation cue span from the participating system did not exist in the reference standard annotations
false negative (FN) = an annotation cue span from the reference standard did not exist in the participating system annotations

Table 1: System Performance, Task 2a: predict each attribute's normalization slot value. Accuracy: overall (official ranking result)

Attribute	System ID ({team}.{system})	Accuracy
Overall Average	TeamHITACHI.2	0.868
	TeamHITACHI.1	0.854
	RelAgent.2	0.843
	RelAgent.1	0.843
	TeamHCMUS.1	0.827
	DFKI-Medical.2	0.822
	LIMSI.1	0.804
	DFKI-Medical.1	0.804
	TeamUEvora.1	0.802
	LIMSI.2	0.801
	ASNLP.1	0.793
	TeamCORAL.1.add	0.790
	TeamGRIUM.1	0.780
	HPI.1	0.769

Recall =

$$\frac{TP}{(TP + FN)} \tag{1}$$

Precision =

$$\frac{TP}{(TP + FP)} \tag{2}$$

F1-score =

$$2\frac{(Recall * Precision)}{(Recall + Precision)} \tag{3}$$

3 Results

Participating teams included between 1-4 people and competed from Canada (team GRIUM), France (team LIMSI), Germany (teams HPI and DFKI-Medical), India (teams RelAgent and HITACHI), Japan (team HITACHI), Portugal (team UEvora), Taiwan (team ASNLP), Vietnam (team HCMUS) and USA (team CORAL). Participants represented academic and industrial institutions including LIMSI-CNRS, University of Alabama at Birmingham, Hasso Plattner Institute, University of Heidelberg, Academia Sinica, DIRO, University of Science, RelAgent Tech Pvt Ltd, University of Evora, Hitachi, International Institute of Information Technology, and German Research Center for AI (DFKI). In total, ten teams submitted systems for Task 2a. Four teams submitted two runs. For Task 2b, three teams submitted systems, one of them submitted two runs.

3.1 System Performance on Task 2a

As shown in Table 1, the HITACHI team system (run 2) had the highest performance in Task 2a, with an overall average accuracy of 0.868. For the individual attributes, team HITACHI had the highest performance for Negation Indicator (0.969), Uncertainty Indicator (0.960), Course Class (0.971), Severity Class (0.982), Conditional Class (0.978), Body Location (0.797) and DocTime Class (0.328), Tables 2 and 3. The HCMUS team had the highest performance for the attribute Subject Class (0.995), and three teams (HPI, RelAgent, Coral) had the highest performance for the attribute Temporal Expression (0.864). For the attribute Generic Class, most teams correctly predicted no change in the normalization value.

3.2 System Performance on Task 2b

For Task 2b, the HITACHI team system (run 2) had the highest performance, with an overall average F1-score (strict) of 0.676 (Table 4). Team HITACHI also had the highest performance (strict) for the individual attributes Negation Indicator (0.913), Uncertainty Indicator (0.9561), Course Class (0.645), Severity Class (0.847), Conditional Class (0.638), Generic Class (0.225) and Body Location (0.854). The HCMUS team had the highest performance for the attribute Subject Class (0.857), and Temporal Expression (0.287).

4 Discussion

We released an extended ShARe corpus through Task 2 of the ShARe/CLEFeHealth Evaluation Lab. This corpus contains disease/disorder templates with ten semantic attributes. In the evaluation lab, we evaluated systems on the task of normalizing semantic attribute values overall and by attribute type (Task 2a), as well as on the task of assigning attribute cue slot values (Task 2b). This is a unique clinical NLP challenge - no previous challenge has targeted such rich semantic annotations. Results show that high overall average accuracy can be achieved by NLP systems on the task of normalizing semantic attribute values, but that performance levels differ greatly between individual attribute types, which was also reflected in the results for cue slot prediction (Task 2b). This corpus and the participating team system results are an important contribution to the research community and the focus on rich semantic information is unprecedented.

Acknowledgments

We greatly appreciate the hard work and feedback of our program committee members. We also want to thank all participating teams. This shared task was partially supported by the CLEF Initiative, the ShARe project funded by the United States National Institutes of Health (R01GM090187), the US Office of the

National Coordinator of Healthcare Technology, Strategic Health IT Advanced Research Projects (SHARP) 90TR0002, and the Swedish Research Council (350-2012-6658).

References

1. Center for Medicare, Medicaid Services: Eligible professional meaningful use menu set measures: Measure 5 of 10. http://www.cms.gov/Regulations-and-Guidance/Legislation/EHRIncentivePrograms/downloads/5_Patient_Electronic_Access.pdf Accessed: 2014-06-16.
2. Eutopian Union: Directive 2011/24/EU of the European Parliament and of the Council of 9 march 2011. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:088:0045:0065:en:PDF Accessed: 2014-06-16.
3. Mowery, D., Jordan, P., Wiebe, J., Harkema, H., Dowling, J., Chapman, W.: Semantic annotation of clinical events for generating a problem list. AMIA Annu Symp Proc (2013) 1032–1041
4. Apostolova, E., Channin, D., Demner-Fushman, D., Furst, J., Lytinen, S., Raicu, D.: Automatic segmentation of clinical texts. Conf Proc IEEE Eng Med Biol Soc (2009) 5905–5908
5. Engel, K., Buckley, B., Forth, V., McCarthy, D., Ellison, E., Schmidt, M., Adams, J.: Patient understanding of emergency department discharge summary instructions: Where are knowledge deficits greatest? Acad Emerg Med **19**(9) (2012) E1035–E1044
6. Uzuner, Ö., Mailoa, J., Ryan, R., Sibanda, T.: Semantic relations for problem-oriented medical records. Artif Intell Med **50**(2) (October 2010) 63–73
7. Sun, W., Rumshisky, A., Uzuner, O.: Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. J Am Med Inform Assoc **20** (2013) 806–813
8. Kelly, L., Goeuriot, L., Suominen, H., Schreck, T., Leroy, G., Mowery, D., Velupillai, S., Martinez, D., Chapman, W., Zuccon, G., Palotti, J.: Overview of the share/clef ehealth evaluation lab 2014. In: Lecture Notes in Computer Science (LNCS). (2014)
9. Suominen, H., Schreck, T., Leroy, G., Hochheiser, H., Goeuriot, L., Kelly, L., Mowery, D., Nualart, J., Ferraro, G., Keim, D.: Task 1 of the CLEF eHealth Evaluation Lab 2014: visual-interactive search and exploration of eHealth data. In Cappellato, L., Ferro, N., Halvey, M., Kraaij, W., eds.: CLEF 2014 Evaluation Labs and Workshop: Online Working Notes, Sheffield, UK, CLEF (2014)
10. Goeuriot, L., Kelly, L., Lee, W., Palotti, J., Pecina, P., Zuccon, G., Hanbury, A., Gareth J.F. Jones, H.M.: ShARe/CLEF eHealth Evaluation Lab 2014, Task 3: User-centred health information retrieval. In Cappellato, L., Ferro, N., Halvey, M., Kraaij, W., eds.: CLEF 2014 Evaluation Labs and Workshop: Online Working Notes, Sheffield, UK, CLEF (2014)
11. Elhadad, N., Chapman, W., OGorman, T., Palmer, M., Savova, G.: The ShARe schema for the syntactic and semantic annotation of clinical texts. under review.
12. : ShARe CLEF eHealth website task 2 information extraction. https://sites.google.com/a/dcu.ie/clefehealth2014/task-2/2014-dataset Accessed: 2014-06-16.
13. : ShARe CLEF eHealth website task 2 information extraction. https://drive.google.com/file/d/0B7oJZ-fwZvH5ZXFRTGl6U3Z6cVE/edit?usp=sharing Accessed: 2014-06-16.

14. Saeed, M., Lieu, C., Raber, G., Mark, R.: MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. Comput Cardiol **29** (2002)
15. CITI: Collaborative Institutional Training Initiative. https://www.citiprogram.org/ Accessed: 2013-06-30.
16. NIH: National Institute of Health - ethics training module. http://ethics.od.nih.gov/Training/AET.htm Accessed: 2013-06-30.
17. Physionet: Physionet site. https:http://www.physionet.org/ Accessed: 2013-06-30.
18. Hripcsak, G., Rothschild, A.: Agreement, the F-measure, and reliability in information retrieval. J Am Med Inform Assoc **12**(3) 296–8

Table 2: System Performance, Task 2a: predict each attribute's normalization slot value. Accuracy per attribute type - Attributes *Negation Indicator*, *Subject Class*, *Uncertainty Indicator*, *Course Class*, *Severity Class*, *Conditional Class*.

Attribute	System ID	Accuracy	Attribute	System ID	Accuracy
Negation Indicator	TeamHITACHI.2	0.969	Subject Class	TeamHCMUS.1	0.995
	RelAgent.2	0.944		TeamHITACHI.2	0.993
	RelAgent.1	0.941		TeamHITACHI.1	0.990
	TeamASNLP	0.923		TeamUEvora.1	0.987
	TeamGRIUM.1	0.922		DFKI-Medical.1	0.985
	TeamHCMUS.1	0.910		DFKI-Medical.2	0.985
	LIMSI.1	0.902		LIMSI.1	0.984
	LIMSI.2	0.902		RelAgent.2	0.984
	TeamUEvora.1	0.901		RelAgent.1	0.984
	TeamHITACHI.1	0.883		LIMSI.2	0.984
	DFKI-Medical.2	0.879		TeamHPI	0.976
	DFKI-Medical.1	0.876		TeamCORAL.1.add	0.926
	TeamCORAL.1.add	0.807		TeamASNLP	0.921
	TeamHPI	0.762		TeamGRIUM.1	0.611
Uncertainty Indicator	TeamHITACHI.1	0.960	Course Class	TeamHITACHI.2	0.971
	RelAgent.2	0.955		TeamHITACHI.1	0.971
	RelAgent.1	0.955		RelAgent.1	0.970
	TeamUEvora.1	0.955		RelAgent.2	0.967
	TeamCORAL.1.add	0.941		TeamGRIUM.1	0.961
	DFKI-Medical.1	0.941		TeamCORAL.1.add	0.961
	DFKI-Medical.2	0.941		TeamASNLP	0.953
	TeamHITACHI.2	0.924		TeamHCMUS.1	0.937
	TeamGRIUM.1	0.923		DFKI-Medical.1	0.932
	TeamASNLP	0.912		DFKI-Medical.2	0.932
	TeamHPI	0.906		TeamHPI	0.899
	TeamHCMUS.1	0.877		TeamUEvora.1	0.859
	LIMSI.1	0.801		LIMSI.1	0.853
	LIMSI.2	0.801		LIMSI.2	0.853
Severity Class	TeamHITACHI.2	0.982	Conditional Class	TeamHITACHI.1	0.978
	TeamHITACHI.1	0.982		TeamUEvora.1	0.975
	RelAgent.2	0.975		RelAgent.2	0.963
	RelAgent.1	0.975		RelAgent.1	0.963
	TeamGRIUM.1	0.969		TeamHITACHI.2	0.954
	TeamHCMUS.1	0.961		TeamGRIUM.1	0.936
	DFKI-Medical.1	0.957		LIMSI.1	0.936
	DFKI-Medical.2	0.957		TeamASNLP	0.936
	TeamCORAL.1.add	0.942		LIMSI.2	0.936
	TeamUEvora.1	0.919		TeamCORAL.1.add	0.936
	TeamHPI	0.914		DFKI-Medical.1	0.936
	TeamASNLP	0.912		DFKI-Medical.2	0.936
	LIMSI.1	0.900		TeamHCMUS.1	0.899
	LIMSI.2	0.900		TeamHPI	0.819

Table 3: System Performance, Task 2a: predict each attribute's normalization slot value. Accuracy per attribute type - Attributes *Generic Class*, *Body Location*, *DocTime Class* and *Temporal Expression*.

Attribute	System ID	Accuracy	Attribute	System ID	Accuracy
Generic Class	TeamGRIUM.1	1.000	Body Location	TeamHITACHI.2	0.797
	LIMSI.1	1.000		TeamHITACHI.1	0.790
	TeamHPI	1.000		RelAgent.2	0.756
	TeamHCMUS.1	1.000		RelAgent.1	0.753
	RelAgent.2	1.000		TeamGRIUM.1	0.635
	TeamASNLP	1.000		DFKI-Medical.2	0.586
	RelAgent.1	1.000		TeamHCMUS.1	0.551
	LIMSI.2	1.000		TeamASNLP	0.546
	TeamUEvora.1	1.000		TeamCORAL.1.add	0.546
	DFKI-Medical.1	1.000		TeamUEvora.1	0.540
	DFKI-Medical.2	1.000		LIMSI.1	0.504
	TeamHITACHI.2	0.990		LIMSI.2	0.504
	TeamCORAL.1.add	0.974		TeamHPI	0.494
	TeamHITACHI.1	0.895		DFKI-Medical.1	0.486
DocTime Class	TeamHITACHI.2	0.328	Temporal Expression	TeamHPI	0.864
	TeamHITACHI.1	0.324		RelAgent.2	0.864
	LIMSI.1	0.322		RelAgent.1	0.864
	LIMSI.2	0.322		TeamCORAL.1.add	0.864
	TeamHCMUS.1	0.306		TeamUEvora.1	0.857
	DFKI-Medical.1	0.179		DFKI-Medical.2	0.849
	DFKI-Medical.2	0.154		LIMSI.1	0.839
	TeamHPI	0.060		TeamHCMUS.1	0.830
	TeamGRIUM.1	0.024		TeamASNLP	0.828
	RelAgent.2	0.024		TeamGRIUM.1	0.824
	RelAgent.1	0.024		LIMSI.2	0.806
	TeamUEvora.1	0.024		TeamHITACHI.2	0.773
	TeamASNLP	0.001		TeamHITACHI.1	0.766
	TeamCORAL.1.add	0.001		DFKI-Medical.1	0.750

Table 4: System Performance, Task 2b: predict each attribute's cue slot value. Strict and Relaxed F1-score, Precision and Recall (overall and per attribute type)

Attribute	System ID	Strict			Relaxed		
		F1-score	Precision	Recall	F1-score	Precision	Recall
Overall Average	TeamHITACHI.2	0.676	0.620	0.743	0.724	0.672	0.784
	TeamHITACHI.1	0.671	0.620	0.731	0.719	0.672	0.773
	TeamHCMUS.1	0.544	0.475	0.635	0.648	0.583	0.729
	HPI.1	0.190	0.184	0.197	0.323	0.314	0.332
Negation Indicator	TeamHITACHI.2	0.913	0.955	0.874	0.926	0.962	0.893
	TeamHITACHI.1	0.888	0.897	0.879	0.905	0.912	0.897
	TeamHCMUS.1	0.772	0.679	0.896	0.817	0.735	0.919
	HPI.1	0.383	0.405	0.363	0.465	0.488	0.444
Subject Class	TeamHCMUS.1	0.857	0.923	0.800	0.936	0.967	0.907
	TeamHITACHI.1	0.125	0.068	0.760	0.165	0.092	0.814
	TeamHITACHI.2	0.112	0.061	0.653	0.152	0.085	0.729
	HPI.1	0.106	0.059	0.520	0.151	0.086	0.620
Uncertainty Indicator	TeamHITACHI.2	0.561	0.496	0.647	0.672	0.612	0.746
	TeamHITACHI.1	0.514	0.693	0.408	0.655	0.802	0.553
	TeamHCMUS.1	0.252	0.169	0.494	0.386	0.275	0.646
	HPI.1	0.166	0.106	0.376	0.306	0.209	0.572
Course Class	TeamHITACHI.1	0.645	0.607	0.689	0.670	0.632	0.712
	TeamHITACHI.2	0.642	0.606	0.682	0.667	0.632	0.705
	TeamHCMUS.1	0.413	0.316	0.594	0.447	0.348	0.628
	HPI.1	0.226	0.153	0.435	0.283	0.196	0.510
Severity Class	TeamHITACHI.2	0.847	0.854	0.839	0.850	0.857	0.843
	TeamHITACHI.1	0.843	0.845	0.841	0.847	0.848	0.845
	TeamHCMUS.1	0.703	0.665	0.746	0.710	0.672	0.752
	HPI.1	0.364	0.306	0.448	0.396	0.336	0.483
Conditional Class	TeamHITACHI.1	0.638	0.744	0.559	0.801	0.869	0.743
	TeamHITACHI.2	0.548	0.478	0.643	0.729	0.669	0.800
	TeamHCMUS.1	0.307	0.225	0.484	0.441	0.340	0.625
	HPI.1	0.100	0.059	0.315	0.317	0.209	0.658
Generic Class	TeamHITACHI.1	0.225	0.239	0.213	0.304	0.320	0.289
	TeamHITACHI.2	0.192	0.385	0.128	0.263	0.484	0.181
	HPI.1	0.100	0.058	0.380	0.139	0.081	0.470
	TeamHCMUS.1	0.000	0.000	0.000	0.000	0.000	0.000
Body Location	TeamHITACHI.2	0.854	0.880	0.829	0.874	0.897	0.853
	TeamHITACHI.1	0.847	0.866	0.829	0.868	0.885	0.852
	TeamHCMUS.1	0.627	0.568	0.700	0.750	0.701	0.807
	HPI.1	0.134	0.298	0.086	0.363	0.611	0.258
Temporal Expression	TeamHCMUS.1	0.287	0.313	0.265	0.354	0.383	0.329
	TeamHITACHI.2	0.275	0.226	0.354	0.370	0.310	0.458
	TeamHITACHI.1	0.269	0.217	0.356	0.364	0.300	0.461
	HPI.1	0.000	0.000	0.000	0.000	0.000	0.000