Digital Personalized Health and Medicine
L.B. Pape-Haugaard et al. (Eds.)

doi:10.3233/SHTI200116

An Evolutionary Approach to the Annotation of Discharge Summaries

Christina LOHR[1ab], Luise MODERSOHN [ab], Johannes HELLRICH [ab], Tobias KOLDITZ [ab], Udo HAHN [ab]
[a] *Jena University Language & Information Engineering (JULIE) Lab Friedrich-Schiller-Universität Jena, Germany*
[b] *SMITH Consortium of the German Medical Informatics Initiative*

Abstract. We here describe the evolution of annotation guidelines for major clinical named entities, namely *Diagnosis*, *Findings* and *Symptoms*, on a corpus of approximately 1,000 German discharge letters. Due to their intrinsic opaqueness and complexity, clinical annotation tasks require continuous guideline tuning, beginning from the initial definition of crucial entities and the subsequent iterative evolution of guidelines based on empirical evidence. We describe rationales for adaptation, with focus on several metrical criteria and task-centered clinical constraints

Keywords. Clinical text corpus, German discharge summaries, Annotation guideline

1. Introduction

The annotation of clinical documents is a challenging task due to the intrinsic opaqueness and complexity of the clinical domain. Clinical experts do not necessarily agree on fundamental judgments relevant for annotation decisions, e.g., the distinction between symptoms, diagnoses and other medically relevant phenomena. Reports of medical annotation campaigns often disregard this lack of conclusiveness insofar as the painstaking adaptive processes underlying the vernier adjustment of annotation guidelines are hidden from the scientific community. Most often only the final outcome of consensus-seeking processes is made publicly available. Annotated clinical text corpora focusing on diagnoses came up with clinical shared tasks, e.g., I2B2 [1], CLEF EHEALTH [2], and SEMEVAL [3]. The annotation of major clinical entities is typically considered as a very hard decision task, with agreement scores (*F*-scores, κ-or α-values) usually in the range between 0.7 and 0.8, even after several modification rounds for annotation guidelines (see Table 1 for an overview of previous annotation campaigns related to clinical documents).

We here want to shed light on the impact of different annotation policies and propose robust metrical decision criteria in order to find out which kinds of adaptive steps have particular impact on annotators' (dis)agreement and how such factors can be controlled by empirical evidence. We illustrate an evolutionary approach to the development of

[1] Corresponding Author: email: christina.lohr@uni-jena.de

annotation guidelines for major clinical named entities, namely *Diagnoses*, *Findings* and *Symptoms*, on a corpus of roughly 1,000 German discharge summaries.

Table 1. Overview of clinical annotation studies, incl. annotated entities and agreement scores (*F*-score, Cohen's κ (κ_c), Siegel and Castellan's κ ($\kappa_{s\&c}$) or Krippendorff's α)

Authors	Data	Annotated Entities	Agreement
English language text corpora			
Wang et al. [4]	300 clinical notes	S_{NOMED}-CT concepts	$F = 0.88$
Savova et al. [5]	273 clinical notes	U_{MLS} semantic groups	$\kappa_c = 0.727$
Doğan et al. [6]	793 PubMed abstracts	diseases	$F \approx 0.9$
Albright et al. [7]	13K pathol. report sentences	U_{MLS} semantic groups	$F \in [0.7, 0.75]$
Hahn et al. [8]	200 Medline abstracts	diseases, symptoms, anatomy	$\kappa_{s\&c} \in [0.7, 0.8]$
Patel et al. [9]	5K different documents	U_{MLS} semantic groups	$\kappa_c \in [0.68, 0.97]$
Multilingual text corpora (including English and German language)			
Miñarro-Giménez et al. [10]	60 short texts	terminology terms S_{NOMED}-CT and U_{MLS}	$\alpha = [0.4, 0.9]$
Kors et al. [11]	800 Medline titles, 500 drug labels, 150 patent claims	biomedical concepts based on U_{MLS} semantic groups	$F = 0.79$
Non-English language text corpora			
Toepfer et al. [12]	140 German ECG reports	fine grained information extraction	$F = 0.955$
Skeppstedt et al. [13]	1,148 Swedish EPR reports	disorders, findings, pharma	$F \in [0.69, 0.88]$
Campillos et al. [14]	500 French clinical notes	entities derived from U_{MLS} semantic groups	$F = 0.793$
He et al. [15]	138 Chinese documents		$F = 0.992$

2. Methods

Data and Annotation Setup. This work was carried out on the Jena part of the 3000PA text corpus [16]—1,106 German discharge summaries from the Jena University Hospital's information system given the following constraints: patients had stayed for at least five days on a ward for internal medicine or in an intensive care unit between 2010 and 2015, and were already deceased. These criteria were approved by the local ethics committee (4639-12/15). The corpus consists of approx. 170K sentences and 1.5M tokens. The documents were annotated (using the *Brat Rapid Annotation Tool*)[2] by eight medical students who all had passed their first medical licensing exam. We ran through four iterations of guideline revisions, tool configuration, team instruction, annotation work, agreement computation, error analysis and subsequent team discussion. This process was supervised by one annotation manager. In each iteration, 5 to 10 documents and 50 in the final round were annotated by each student (for agreement values see Table 2).

Evaluation Measures. A challenging problem for medical annotations is posed by determining the proper text span of named entities. We thus calculate inter-annotator agreement (IAA) by using the pair-wise average *F*-score [17] of instances and tokens. An *instance* is a single composite annotation unit that consists of one or more tokens (e.g., *"renal insufficiency"* denotes an instance with two tokens, *"renal"* and *"insufficiency"*). Instance agreement is a strong criterion, as annotations with only slightly differing spans

[2] https://brat.nlplab.org/

count as non-matching. Token-based agreement describes the overlap of individual tokens for annotations of the same instance type[3].

Table 2. Inter-annotator agreement (IAA; pair-wise average F-score incl. standard deviation (σ)) of training iterations and the final annotation round (incl. the number of documents being used per iteration (in brackets)), for each document over all eight annotators w.r.t. instances (inst.) and tokens (tok.)

	Iteration 1 (10)		**Iteration 2** (5)		**Iteration 3** (10)		**Iteration 4** (10)		**Final** (50)	
Category	**inst.**	**tok.**	**inst.**	**tok.**	**inst.**	**tok.**	**inst.**	**tok.**	**inst.**	**tok.**
Diagnosis	.504	.633	.534	.630	.593	.705	.614	.688	0.637	0.758
Findings	.354	.545	.698	.896	.682	.887	.628	.864	0.684	0.860
Symptoms	.583	.617	.441	.657	.424	.729	.532	.669	0.611	0.718
Anatom. Loc.	.455	.624	.735	.876	–	–	–	–	–	–
Procedures	.252	.301	.527	.580	–	–	–	–	–	–
Avg. F-score	**.410**	**.563**	**.671**	**.832**	**.653**	**.873**	**.621**	**.837**	**.648**	**.839**
σ (F-score)	.079	.076	.056	.030	.064	.027	.065	.036	.056	.032

Annotation Schema. We started our annotation campaign with five types of medical entities: *Diagnosis* as *diseases* as listed in the ICD-10 (excluding symptom chapter *'R'*); *Findings* as the summary of all information about a patient's medical status reported by medical staff based on the results from physical examination or the use of medical devices (e.g., x-ray); *Symptoms* as the summary of all health affecting information and complaints reported by the patients or their relatives; *Anatomical Location* as the summary of all anatomical parts of the body excluding body fluids and body secretions; *Procedures* as the summary of all medical interventions like surgery, non-surgical therapeutic measures, etc. The administration of (new) medication belongs to the latter category, but was excluded here since medication had already been annotated elsewhere [16].

Iterative Process. The average agreement value (*F*-score) of the first iteration reached .410 on instances and .563 on tokens which is clearly below expectations. During the second iteration *Anatomical Locations* were automatically pre-annotated based on the German part of the *Foundational Model of Anatomy (FMA)* of the UMLS using the JUFIT tool. This step increased the IAA for *Anatomical Location* up to .735 on instances and .876 on tokens. Still, it was removed from further annotation rounds because of its overlap with mentions of *Diagnoses*, *Findings* or *Symptoms*. Also discussions with the annotators pointed at a painfully large number of single decisions and double annotations to be made resulting in an average annotation time of one hour per document. In particular, distinguishing *Procedures* from *Medications* (e.g., *"long-term oxygen therapy"*) and *Diagnoses* was considered a difficult task—a previous intervention can be coded as *Procedures* and *Diagnoses* (e.g., *Z98.8—Other specified postsurgical states*). As a clear criterion to minimize the overlap between *Diagnoses* and *Procedures* could not be found we eliminated *Procedures* from the current annotation project and will deal with it separately. Following the approach of Hahn et al. [8] for annotating long and short text spans, we required the attribute *Complexity* to be tagged for all entities containing more than one piece of annotation-relevant information (e.g. *"Pupils middle and isocor"*). *Diagnosis*, *Symptoms* and *Findings* were assigned two attributes: *Time* with the value set *previous*, *recurrent*, *uncertain* and *Modality* with the value set *suspected*, *excluded*, *uncertain.* Negations or vague descriptions were defined to be part of the long annotation span.

[3] We used TREETAGGER, accessible via https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Baseline Classifier. We trained a (CRF-based) JCoRe pipeline for named entity recognition of *Diagnosis*, *Findings* and *Symptoms*; evaluation with 10-fold cross-validation.

Table 3. Instances and tokens of entity types *Diagnosis*, *Findings*, *Symptoms*, with the percentage distribution of their attributes *Complexity*, *Modality*, *Time*; baseline classifier (BC) results (avg. token distribution depending on all of the 1.863M tokens; all other percentage values depending on the total number of entity instances)

	Inst.	Tokens		Avg. span	Com-plexity	Modality			Time			BC
						sus.	exc.	unc.	rec.	unc.	pre.	
Entity type	freq.	freq.	%	tok.	%	%	%	%	%	%	%	F
Diagnosis	55K	123K	6.6	2.2	3.46	**4.44**	**9.31**	1.49	0.67	0.01	**6.66**	.475
Findings	150K	**663K**	**35.6**	4.4	**17.15**	1.24	**17.36**	1.40	0.13	0.02	1.81	.805
Symptom	8K	26K	1.4	3.3	**14.21**	0.10	**22.16**	0.13	0.96	–	1.57	.143
Σ Annotations	163K	**812K**	**43.40**	3.8	13.50	2.02	**15.45**	1.38	0.30	0.01	3.05	–

3. Results

The final average IAA values were .648 on instances and .839 on tokens (see Table 2). Table 3 illustrates the distribution of the different entities and their attributes. About one third of the entire corpus consists of *Findings*. *Diagnoses* added up to around 6.6%, whereas *Symptoms* were rare with less than 2%. Around 13% of all annotations were marked as *complex*; roughly every fifth instance of *Findings* and *Symptoms* contained more than one information unit. Compared to *Findings* and *Symptoms*, *Diagnoses* were more precisely described. 10–20% of all annotations were *negated*. The majority of negated entities were *Symptoms* and *Findings* (≈20%). The attributes *suspected* and *uncertain* were rarely used, only about 5% of *Diagnosis* annotations were marked as *suspected*. Temporal annotations of *Findings* and *Symptoms* were hardly used, ≈ 7% of *Diagnoses* were *previous*. The average *F*-score results of our (admittedly simple) CRF-based classifier are .805 for *Findings*, .475 for *Diagnosis* and .143 for *Symptoms*.

4. Discussion

The final IAA values (0.7–0.8) could be improved and are comparable to, if not higher than, annotation campaigns covering the listed entity types in languages other than German (cf. Table 1). Category overlap or hard to disentangle semantic interdependencies must be resolved for the sake of acceptable IAA values—either by generalizing cate- gory definitions, by stretching the granularity of annotation spans, by adding complexity markers (thus delegating finer distinctions to future annotation studies), or, in the worst case, by eliminating sloppy named entity types from the current annotation process.

5. Conclusion

We presented an evolutionary approach to an annotation scheme for *Diagnoses*, *Findings* and *Symptoms* and its application to German discharge summaries. We struggled with

criteria that control the number of iterations and the directions of adaptations. Generalizing from the our annotation task, we found that IAA is a solid metrical indicator for task complexity or inconclusive annotation guidelines. Both parameters have to be tuned in case of too low IAA values. Annotation time is a metrical indicator for task complexity.

Acknowledgements Work supported by BMBF within the SMITH project under grant 01ZZ1803G and DFG under grant HA 2079/8-1 within the STAKI2B2 project. We thank all annotators, as well as André Scherag, Danny Ammon, and all members of the Data Integration Center of the Jena University Hospital.

References

[1] Uzuner O, South BR, Shen S, and DuVall SL, "2010 i2B2/VA Challenge on concepts, assertions, and relations in clinical text," *JAMIA*, vol. 18, no. 5, pp. 552–556, 2011.
[2] Suominen H, Zhou L, Hanlen L, and Ferraro G, "Benchmarking clinical speech recognition and information extraction: new data, methods, and evaluations," *JMI*, vol. 3, no. 2, p. e19, 2015.
[3] Elhadad N, Pradhan SS, Lipsky Gorman S, Manandhar S, Chapman WW, and Savova GK, "SemEval-2015 Task 14: Analysis of Clinical Text," in *SemEval 2015*, pp. 303–310, 2015.
[4] Wang Y, "Annotating and recognising named entities in clinical notes," in *Proceedings of the Student Research Workshop @ ACL-IJCNLP 2009*, pp. 18–26, 2009.
[5] Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, and Chute CG, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," *JAMIA*, vol. 17, no. 5, pp. 507–513, 2010.
[6] Doğan RI, Leaman R, and Lu Z, "NCBI disease corpus: a resource for disease name recognition and concept normalization," *Journal of Biomedical Informatics*, vol. 47, pp. 1–10, 2014.
[7] Albright D, Lanfranchi A, Fredriksen A, Styler I, William F, Warner C, Hwang JD, Choi JD, Dligach D, Nielsen RD, Martin J, Ward W, Palmer M, and Savova GK, "Towards comprehensive syntactic and semantic annotations of the clinical narrative," *JAMIA*, vol. 20, pp. 922–930, 012013.
[8] Hahn U, Beisswanger E, Buyko E., Faessler E, Traumüller J, Schröder S, and Hornbostel K, "Iterative refinement and quality checking of annotation guidelines: How to deal effectively with semantically sloppy named entity types, such as pathological phenomena," in *LREC 2012*, pp. 3881–3885, 2012.
[9] Patel P, Davey D, Panchal V, and Pathak P, "Annotation of a large clinical entity corpus," in *EMNLP 2018*, pp. 2033–2042, 2018.
[10] Miñarro-Giménez JA, Cornet R, Jaulent MC, Dewenter H, Thun S, Rosenbeck Gøeg K, Karlsson D, and Schulz S, "Quantitative analysis of manual annotation of clinical text samples," *International Journal of Medical Informatics*, vol. 123, pp. 37–48, 2019.
[11] Kors JA, Clematide S, Akhondi SA, van Mulligen EM, and Rebholz-Schuhmann D, "A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC," *Journal of the American Medical Informatics Association*, vol. 22, no. 5, pp. 948–956, 2015.
[12] Toepfer M, Corovic H, Fette G, Klügl P, Störk S, and Puppe F, "Fine-grained information extraction from German transthoracic echocardiography reports," *BMC MIDM*, vol. 15, p. #91, 2015.
[13] Skeppstedt M, Kvist M, Nilsson GH, and Dalianis H, "Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study," *Journal of Biomedical Informatics*, vol. 49, pp. 148–158, June 2014.
[14] Campillos L, Deléger L, Grouin C, Hamon T, Ligozat AL, and Névéol A, "A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI annOtated Text corpus (Merlot)," *Language Resources and Evaluation*, vol. 52, pp. 571–601, 2018.
[15] He B, Dong B, Guan Y, Yang J, Jiang Z, Yu Q, Cheng J, and Qu C, "Building a comprehensive syntactic and semantic corpus of Chineseclinical texts," *JBI*, vol. 69, pp. 203–217, May 2017.
[16] Hahn U, Matthies F, Lohr C, and Löffler M, "3000PA: Towards a national reference corpus of German clinical language," in *MIE 2018*, no. 247, pp. 26–30.
[17] Hripcsak G and Rothschild AS, "Agreement, the f-measure, and reliability in information retrieval," Journal of the American Medical Informatics Association, vol. 12, no. 3, pp. 296–298, 2005.