MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231029

Classification of Diagnostic Certainty in Radiology Reports with Deep Learning

Kento SUGIMOTO[a,1], Shoya WADA[a,b], Shozo KONISHI[a], Katsuki OKADA[a], Shirou MANABE[a,b], Yasushi MATSUMURA[a,c], Toshihiro TAKEDA[a]
[a] *Department of Medical Informatics, Osaka University Graduate School of Medicine, Osaka, Japan*
[b] *Department of Transformative System for Medical Information, Osaka University Graduate School of Medicine, Osaka, Japan*
[c] *National Hospital Organization Osaka National Hospital, Osaka, Japan*
ORCiD ID: Kento Sugimoto https://orcid.org/0000-0002-6874-2399

Abstract. A radiology report is prepared for communicating clinical information about observed abnormal structures and clinically important findings with referring clinicians. However, such observations and findings are often accompanied by ambiguous expressions, which can prevent clinicians from accurately interpreting the content of reports. To systematically assess the degree of diagnostic certainty for each observation and finding in a report, we defined an ordinal scale comprising five classes: definite, likely, may represent, unlikely, and denial. Furthermore, we applied a deep learning classification model to determine its applicability to in-house radiology reports. We trained and evaluated the model using 540 in-house chest computed tomography reports. The deep learning model achieved a micro F1-score of 97.61%, which indicated that our ordinal scale was suitable for measuring the diagnostic certainty of observations and findings in a report.

Keywords. Diagnostic certainty, radiology report, deep learning

1. Introduction

A radiology report includes observed abnormal structures (hereinafter called *observation*) and clinically important findings (hereinafter called *clinical finding*). To convey the degree of diagnostic certainty of observations and clinical findings, radiologists often use various ambiguous expressions. For example, the diagnostic certainty of "*lung cancer*" could differ from "suggestive of *lung cancer*" to "more likely to be an inflammatory change rather than *lung cancer*." The difference of expressions can cause misinterpretations of the content of reports by clinicians [1–3]. This ambiguity also affects the secondary use of reports. To build a curated dataset from reports, observations and clinical findings with negated or uncertain should be ruled out.

Here, we present a fine-grained ordinal scale to measure the degree of diagnostic certainty of observations and clinical findings in reports. Furthermore, we developed a deep learning classification algorithm and evaluated its performance to verify its applicability to in-house radiology reports.

[1] Corresponding Author: Kento Sugimoto, email: sugimoto.kento@hp-info.med.osaka-u.ac.jp.

2. Methods

2.1. Certainty scale

Our certainty scale mainly comprises three classes: *definite*, *uncertain*, and *denial*. A definite class is assigned to observations and clinical findings that are reported without any ambiguity such as "the *nodule* is present" and "consistent with *lung cancer*." Conversely, a denial class is used for observations and clinical findings that negated certain diagnostic possibilities such as "no evidence of *lung cancer*." An uncertain class is categorized into more fine-grained classes: *likely*, *may represent*, and *unlikely*. We referred to the diagnostic certainty scale developed by Shinagare et al. [4], which was categorized into five classes: *most likely*, *likely*, *may represent*, *unlikely*, and *very unlikely*. For annotation simplicity, we considered the most likely class to be the same as the definite class. Similarly, the very unlikely class was also considered to be the same as the denial class. Thus, five classes were defined.

2.2. Corpus development

This study was approved by the institutional review board of Osaka University Hospital (approval number: 19276). Chest computed tomography reports from 2010 to 2018 stored in the radiology information system at Osaka University Hospital were used. The dataset consisted of 118,078 reports written in Japanese, of which 540 reports were randomly selected for training the deep learning model and evaluating its performance.

To build a gold standard dataset, three medical students performed the annotation process. Annotators were given reports with highlighted terms of observations and clinical findings in each report. Then, they annotated the predefined certainty class to each highlighted term. To achieve consistent annotation, the annotators were provided with a guideline describing the classification criteria of the certainty scale. Annotation disagreements were resolved by a majority vote. Clinicians resolved disagreements among the three annotators if the gold standard could not be determined by a majority vote. In the 540 reports, a total of 4,485 observation and clinical finding terms were annotated. The Fleiss' kappa score [5] to measure inter-annotator agreement was 89.9%, which denotes very high agreement [6]. Table 1 shows the number of each certainty scale class in the dataset.

Table 1. The number of each certainty scale class in the dataset.

Definite	Likely	May represent	Unlikely	Denial	Total
2,242	538	286	112	1,307	4,485

2.3. Classification model

An overview of our system is shown in Figure 1. Our system contains two deep learning components: (1) to extract observation and clinical finding terms according to predefined entities; (2) to classify the certainty scale class of each observation and clinical finding term.

We have previously reported the deep learning algorithm to extract observation and clinical finding terms from reports [7]. The micro F1-scores of our best-performing

model for extracting observation and clinical finding terms were 94.22% and 95.61%, respectively. Additional details have been reported in our previous paper [7].

Next, to determine the certainty scale class of each observation and clinical finding term, we applied the BERT [8] as a classification model, and fed a report into a model. More specifically, given a report with a span of target term, the model predicts the certainty scale class of the term. Observation and clinical finding terms in a report were fed into the model one by one. To identify a span of target term in a report, we inserted entity marker tokens both before and after the target term (Figure 1).

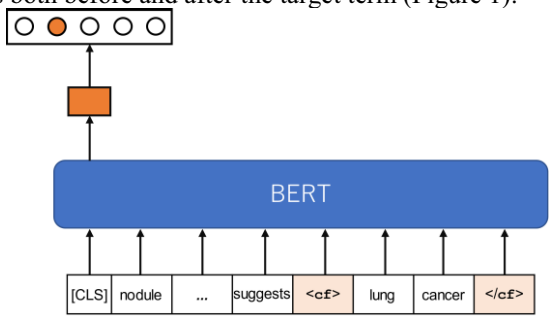

Figure 1. An overview of the certainty scale classification model. A classification model predicts the certainty scale of each extracted term. Entity marker tokens (<cf>, </cf>) are inserted to identify the target term (cf: clinical finding). In this example, the model predicts the certainty scale of the term "lung cancer".

2.4. Evaluation metrics

Strict and relaxed metrics were used in our experiment. A strict metric is a general classification metric that counts as correct when the prediction class is the same as the ground truth class. However, because a strict metric ignores the distance between classes (e.g., likely and may represent vs definite and denial), we cannot evaluate the performance of the ordinal scale in detail. To solve this problem, we introduced a relaxed metric that allows the difference of the nearest neighbor class. As an exception, since negation detection is essential in clinical applications, the denial class is counted as correct only when both the prediction and ground truth are denial. The F1-score was used for both the strict and relaxed metrics. To compare the difference between the strict and relaxed metrics, we also showed an error rate per class, which is 1 − F1-score.

3. Results

A total of 540 annotated reports were divided into 378 reports for training, 54 reports for development, and 108 reports for testing. The best hyperparameter settings were chosen using a development dataset.

Our experimental result is shown in Table 2. Our certainty classification model obtained 97.61% in the strict micro F1-score. Although the entire performance achieved satisfactory results, the may represent and unlikely classes had lower F1-score in the strict metric than other classes. The denial class obtained a strict F1-score of 98.89%, indicating the high performance as a negation detection module. The entire performance in the relaxed metric was 98.91%, and all classes yielded higher F1-scores than the strict metric. Error rates in the relaxed metric were under 3% in all classes.

Table 2. Performance metrics of each certainty class.

Class	strict		relaxed	
	F1-score	Error rate	F1-score	Error rate
Definite	98.56%	1.44%	99.01%	0.99%
Likely	96.33%	3.67%	100.00%	0.00%
May represent	91.55%	8.45%	97.30%	2.70%
Unlikely	87.80%	12.20%	97.30%	2.70%
Denial	98.89%	1.11%	98.89%	1.11%
Total	**97.61%**	**2.39%**	**98.91%**	**1.09%**

4. Discussion

While the definite and denial classes achieved satisfactory results, the may represent and unlikely classes had lower F1-scores in the strict metric. The small size of these classes in training data probably hindered learning. Table 2 shows that the error rate in the relaxed metric is less than half the error rate in the strict metric, which indicates that a lot of the discrepancy is between the nearest neighbor class. A confusion matrix for each certainty scale class is plotted (Figure 2). This reveals that our model predicts some observations and findings as the denial class even though the ground truth was the definite class and vice versa. Some error examples are shown in Table 3. In Example 1, due to the complex text structure containing the negation word, the model misclassified the case as denial. In Example 2, although the words "*almost disappeared*" implicitly negated the observation, the medical students annotated it as definite. Similarly, "*unclear*" in Example 3 is not a word that positively indicates the existence of observations or findings. In cases such as Examples 2 and 3, we believe that the predicted class is probably reasonable. Similarly, the predicted class in Example 4 is probably more reasonable than the gold standard one. These results reveal the robustness of our model against the noise of training data.

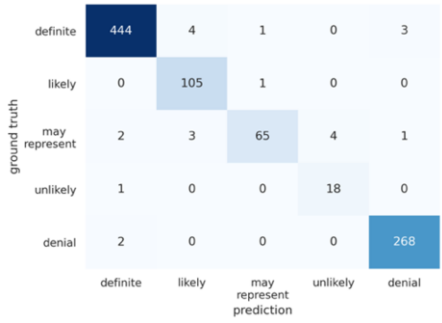

Figure 2. Confusion matrix for each certainty class.

Table 3. Examples of the discrepancy between ground truth and model prediction.

No	Report	Ground Truth	Prediction
Example 1	The margin of the mass lesion is **not** clear ...	Definite	Denial
Example 2	Ground glass opacity **almost disappeared** ...	Definite	Denial
Example 3	Cystic lesion is **unclear** ...	Definite	Denial
Example 4	Edema was **improved** ...	Denial	Definite

One limitation of this study is generalizability. We only trained and evaluated the model using reports collected from a single institution. To ensure generalizability, studies on datasets from outside our institution would be needed.

5. Conclusions

We presented an ordinal scale to measure the degree of diagnostic certainty. The deep learning model achieved satisfactory results, which demonstrated that our certainty scale was sufficiently applicable to in-house radiology reports. We believe that this automated classification system will be helpful to clinicians to reduce misinterpretation of radiology reports and contribute to building a curated dataset for secondary use.

Acknowledgments

This work was supported by the Council for Science, Technology and Innovation (CSTI), cross-ministerial Strategic Innovation Promotion Program (SIP), "Innovative AI Hospital System" (Funding Agency: National Institute of Biomedical Innovation, Health and Nutrition (NIBIOHN).

References

[1] Clinger NJ, Hunter TB, Hillman BJ. Radiology reporting: attitudes of referring physicians. Radiology. 1988 Dec;169(3):825-6, doi: 10.1148/radiology.169.3.3187005.

[2] Khorasani R, Bates DW, Teeger S, Rothschild JM, Adams DF, Seltzer SE. Is terminology used effectively to convey diagnostic certainty in radiology reports? Acad Radiol. 2003 Jun;10(6):685–8, doi: 10.1016/s1076-6332(03)80089-2.

[3] Rosenkrantz AB, Kiritsy M, Kim S. How "consistent" is "consistent"? A clinician-based assessment of the reliability of expressions used by radiologists to communicate diagnostic confidence. Clin Radiol. 2014 Jul;69(7):745–9, doi: 10.1016/j.crad.2014.03.004.

[4] Shinagare AB, Alper DP, Hashemi SR, Chai JL, Hammer MM, Boland GW, Khorasani R. Early Adoption of a Certainty Scale to Improve Diagnostic Certainty Communication. J Am Coll Radiol. 2020 Oct;17(10):1276–84, doi: 10.1016/j.jacr.2020.03.033.

[5] Fleiss JL. Measuring nominal scale agreement among many raters. Psychol Bull. 1971 Nov;76(5):378–82, doi: 10.1037/h0031619.

[6] Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977 Mar;33(1):159–74, doi: 10.2307/2529310.

[7] Sugimoto K, Takeda T, Oh J-H, Wada S, Konishi S, Yamahata A, Manabe S, Tomiyama N, Matsunaga T, Nakanishi K, Matsumura Y. Extracting clinical terms from radiology reports with deep learning. J Biomed Inform. 2021 Apr 1;116:103729, doi: 10.1016/j.jbi.2021.103729.

[8] Devlin J, Chang M-W, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, Minnesota: Association for Computational Linguistics; 2019 May. p. 4171–86, doi: 10.18653/v1/N19-1423.