# Detecting Insertion, Substitution & Deletion Errors in Radiology Reports Using Sequence-to-Sequence Models

*John R. Zech, M.D., M.A.[1], Jessica Z. Forde, M.A.[2], Joseph J. Titano, M.D.[1], Deepak A. Kaji, B.A.[1], Anthony B. Costa, Ph.D.[1], Eric Karl Oermann, M.D.[1]*

*[1]Icahn School of Medicine at Mount Sinai, [2]Project Jupyter*

**Background.** Errors in grammar, spelling, and usage in radiology reports are common: a 2015 review at the Mayo Clinic found error rates as high as 19.7% in neuroradiology and as low as 3.2% in chest x-rays, with a 9.7% overall rate [1]. Nearly 20% of errors in the study were clinically material. To detect erroneous insertions, substitutions, and deletions of words in radiology reports, we proposed using neural sequence-to-sequence (seq2seq) models [2].

**Methods.** Head CT and chest x-ray reports from Mt. Sinai Hospital (MSH: n=61,722 and 818,978, respectively), Mt. Sinai Queens (MSQ: n=30,145 and 194,309) and MIMIC-III (n=32,259 and 54,685) were preprocessed and separated into sentences such that each report was included in either the training (80%), tune (10%), or test (10%) set. Corruptions of insertion, substitution, and deletion were introduced, each with probability 1% for each word in an original sentence; insertions ranged between 1-4 words with equal probability [3]. This process was repeated 5x for head CTs and 2x for chest x-rays on training data and once on tune and test data. Three seq2seq models were trained using head CTs from MSH, chest x-rays from MSH, and head CTs from all three sites. Each model was trained with corrupted sentences to predict original, uncorrupted sentences with OpenNMT-py using a bidirectional LSTM with 512 hidden units, 2 layers, and Luong attention [4]. Models were trained using SGD with learning rate decay until tune perplexity plateaued for 3 epochs. Single-modality (head CT or chest x-ray only) model performance was assessed for the same modality across sites and across modalities at MSH. The jointly trained head CT model was assessed at each site. We measured the fraction of test sentences for which seq2seq correctly predicted a sentence different from the input. A sample of uncorrupted test sentences was also manually reviewed to assess performance on real-world radiologist errors.

**Results.** In same-site, same-modality test sets, seq2seq detected 90.3% of corrupted sentences in head CTs and 88.2% in chest x-rays with 97.7% and 98.8% specificity, respectively. For errors introduced in same-site, same-modality test sets, seq2seq recovered 74.1% of the original head CT sentences and 73.0% of the chest x-ray sentences (Table 1). When it did not recover the original sentence, it sometimes offered reasonable alternatives. Seq2seq performance degraded when test sites and modalities were not included in training but performed well when trained jointly on multiple sites. Performance was strongest in the same-site, same-modality comparison as radiology reports are typically standardized within a site but vary between sites, and the language of reports is highly specific to modality. Manual review of original, uncorrupted Mt. Sinai head CT sentences (400 seq2seq-changed and 800 unchanged) demonstrated model PPV 0.393 (157/400; 95% C.I. 0.346-0.441) and NPV 0.986 (789/800; 95% C.I. 0.976-0.992) for detecting sentences containing real-world radiologist errors.

**Conclusion.** Sequence-to-sequence models can effectively detect insertions, substitutions, and deletions of words in radiology reports. To achieve high performance, they require site- and modality-specific training examples.

| | Corrupted | Seq2seq | Original |
|---|---|---|---|
| **Insertion** | the patient is status **sinuses are hypoplastic the** post bilateral cataract surgery | the patient is status post bilateral cataract surgery | the patient is status post bilateral cataract surgery |
| **Substitution** | **perirolandic** white matter ischemic changes | **chronic** white matter ischemic changes | **periventricular** white matter ischemic changes |
| **Deletion** | head ct: there is no evidence of acute intra or extra axial, midline shift or mass effect | head ct: there is no evidence of acute intra or extra axial **hemorrhage**, midline shift or mass effect | head ct: there is no evidence of acute intra or extra axial **hemorrhage**, midline shift or mass effect |

**Table 1:** Examples of corrupted sentences, seq2seq predictions, and original correct sentences for each type of error. Orange indicates an erroneous change from the original sentence, and blue represents an accurate correction.

[1] Ringler MD, Goss BC, and Bartholmai BJ. Syntactic and semantic errors in radiology reports associated with speech recognition software. Health Informatics J., 23(1):3–13, March 2017.

[2] Bahdanau D, Cho K, and Bengio Y. Neural machine translation by jointly learning to align and translate. ICLR 2015.

[3] Paino AT. deep-text-corrector. https://github.com/atpaino/deep-text-corrector.

[4] Klein G, Kim Y, Deng Y, Senellart J, and Rush A. OpenNMT: Open-source toolkit for neural machine translation. ACL 2017.