
Extractive Summarization of EHR Discharge Notes

Emily Alsentzer
Anne Kim

EMILYA@MIT.EDU
ANNEYKIM@MIT.EDU

Abstract

Patient summarization is essential for clinicians to provide coordinated care and practice effective communication. Automated summarization has the potential to save time, standardize notes, aid clinical decision making, and reduce medical errors. Here we provide an upper bound on extractive summarization of discharge notes and develop an LSTM model to sequentially label topics of history of present illness notes. We achieve an F1 score of 0.876, which indicates that this model can be employed to create a dataset for evaluation of extractive summarization methods.

1. Introduction

Summarization of patient information is essential to the practice of medicine. Clinicians must synthesize information from diverse data sources to communicate with colleagues and provide coordinated care. Examples of clinical summarization are abundant in practice; patient handoff summaries facilitate provider shift change, progress notes provide a daily status update for a patient, oral case presentations enable transfer of information from overnight admission to the care team and attending, and discharge summaries provide information about a patient's hospital visit to their primary care physician and other outpatient providers (Feblowitz et al., 2011).

Informal, unstructured, or poor quality summaries can lead to communication failures and even medical errors, yet clinical instruction on how to formulate clinical summaries is ad hoc and informal. Non-existent or limited search functionality, fragmented data sources, and limited visualizations in electronic health records (EHRs) make summarization challenging for providers (Christensen & Grimsno, 2008; Singh et al., 2013; Natarajan et al., 2010). Furthermore, while dictation of EHR notes

allows clinicians to more efficiently document information at the point of care, the stream of consciousness-like writing can hinder the readability of notes. Kripalani et al. show that discharge summaries are often lacking key information, including treatment progression and follow-up protocols, which can hinder communication between hospital and community based clinicians (Kripalani et al., 2007). Recently, St. Thomas Hospital in Nashville, TN stipulated that discharge notes be written within 48 hours of discharge following incidences where improper care was given to readmitted patients because the discharge summary for the previous admission was not completed (Alsentzer).

Automated summary generation has the potential to save clinician time, avoid medical errors, and aid clinical decision making. By organizing and synthesizing a patient's salient medical history, algorithms for patient summarization can enable better communication and care, particularly for chronically ill patients, whose medical records often contain hundreds of notes. In this work, we explore the automatic summarization of discharge summary notes, which are critical to ensuring continuity of care after hospitalization. We (1) provide an upper bound on extractive summarization by assessing how much information in the discharge note can be found in the rest of the patient's EHR notes and (2) develop a classifier for labeling the topics of history of present illness notes, a narrative section in the discharge summary that describes the patient's prior history and current symptoms. Such a classifier can be used to create topic specific evaluation sets for methods that perform extractive summarization. These aims are critical steps in ultimately developing methods that can automate discharge summary creation.

2. Related Work

In the broader field of summarization, automization was meant to standardize output while also saving time and effort. Pioneering strategies in summarization started by extracting "significant" sentences in the whole corpus to build an abstract where "significant" sentences were defined by the number of frequently occurring words (Luhn, 1958). These initial methods did not consider

word meaning or syntax at either the sentence or paragraph level, which made them crude at best. More advanced extractive heuristics like topic modeling (Allahyari & Kochut, 2016), cue word dictionary approaches (Edmundson, 1969), and title methods (Ferreira et al., 2013) for scoring content in a sentence followed soon after. For example, topic modeling extends initial frequency methods by assigning topics scores by frequency of topic signatures, clustering sentences with similar topics, and finally extracting the centroid sentence, which is considered the most representative sentence (Allahyari et al., 2017). Recently, abstractive summarization approaches using sequence-to-sequence methods have been developed to generate new text that synthesizes original text (Paulus et al., 2017; Nallapati et al., 2016; Liu & Pan, 2016; Rush et al., 2015); however, the field of abstractive summarization is quite young.

Existing approaches within the field of electronic health record summarization have largely been extractive and indicative, meaning that summaries point to important pieces in the original text rather than replacing the original text altogether. Few approaches have been deployed in practice, and even fewer have demonstrated impact on quality of care and outcomes (Pivovarov, 2016). Summarization strategies have ranged from extraction of relevant sentences from the original text to form the summary (Moen et al., 2016), topic modeling of EHR notes using Latent Dirichlet allocation (LDA) or bayesian networks (Pivovarov, 2016), and knowledge based heuristic systems (Goldstein & Shahar, 2016). To our knowledge, there is no literature to date on extractive or abstractive EHR summarization using neural networks.

3. Methods

3.1. Data

MIMIC-III is a freely available, deidentified database containing electronic health records of patients admitted to an Intensive Care Unit (ICU) at Beth Israel Deaconess Medical Center between 2001 and 2012. The database contains all of the notes associated with each patient’s time spent in the ICU as well as 55,177 discharge reports and 4,475 discharge addendums for 41,127 distinct patients. Only the original discharge reports were included in our analyses. Each discharge summary was divided into sections (Date of Birth, Sex, Chief Complaint, Major Surgical or Invasive Procedure, History of Present Illness, etc.) using a regular expression.

3.2. Upper Bound on Summarization

Extractive summarization of discharge summaries relies on the assumption that the information in the discharge

summary is documented elsewhere in the rest of the patient’s notes. However, sometimes clinicians will document information in the discharge summary that may have been discussed throughout the hospital visit, but was never documented in the EHR. Thus, our first aim was to determine the upper bound of extractive summarization.

For each patient, we compared the text of the discharge summary to the remaining notes for the patient’s current admission as well as their entire medical record. Concept Unique Identifiers (CUIs) from the Unified Medical Language System (UMLS) were compared in order to assess whether clinically relevant concepts in the discharge summary could be located in the remaining notes (NLM, 2009). CUIs were extracted using Apache cTAKES (Savova et al., 2010) and filtered by removing the CUIs that are already subsumed by a longer spanning CUI. For example, CUIs for “head” and “ache” were removed if a CUI existed for “head ache” in order to extract the most clinically relevant CUIs.

In order to understand which sections of the discharge summaries would be the easiest or most difficult to summarize, we performed the same CUI overlap comparison for the chief complaint, major surgical or invasive procedure, discharge medication, and history of present illness sections of the discharge note separately. We calculated which fraction of the CUIs in each section were located in the rest of the patient’s note for a specific hospital stay. We also calculated what percent of the genders recorded in the discharge summary were also recorded in the structured data for the patient.

3.3. Labeling History of Present Illness Notes

3.3.1. ANNOTATION

We developed a classifier to label topics in the history of present illness (HPI) notes, including demographics, diagnosis history, and symptoms/signs, among others. A random sample of 515 history of present illness notes was taken, and each of the notes was manually annotated by one of eight annotators using the software Multi-document Annotation Environment (MAE) (Rim, 2016). MAE provides an interactive GUI for annotators and exports the results of each annotation as an XML file with text spans and their associated labels for additional processing. 40% of the HPI notes were labeled by clinicians and 60% by non-clinicians. Table 1 shows the instructions given to the annotators for each of the 10 labels. The entire HPI note was labeled with one of the labels, and instructions were given to label each clause in a sentence with the same label when possible.

Table 1. HPI Categories and Annotation Instructions

Demographics	Age, gender, race/ethnicity, language, social history etc.
DiagnosisHistory	Diagnoses prior to current symptoms
MedicationHistory	Medications or treatment prior to current symptoms
ProcedureHistory	Procedures prior to current symptoms
Symptoms/Signs	Current Signs/Symptoms; chief complaint
Vitals/Labs	Any vitals or labs related to current care episode
Procedures/Results	Procedures including CT, CXR, etc. and their results
Meds/Treatments	Medication or treatment administered in current care episode
Movement	Description of patient admitted/transferred between hospitals or care units
Other	Does not fit into any other category

3.3.2. MODEL

Our LSTM model was adopted from prior work by Dernoncourt et al (Dernoncourt et al., 2016). Whereas the Dernoncourt model jointly classified each sentence in a medical abstract, here we jointly classify each word in the HPI summary. Our model consists of four layers: a token embedding layer, a word contextual representation layer, a label scoring layer, and a label sequence optimization layer (Figure 1).

In the following descriptions, lowercase italics is used to denote scalars, lowercase bold is used to denote vectors, and uppercase italics is used to denote matrices.

Token Embedding Layer: In the token embedding layer, pretrained word embeddings are combined with learned character embeddings to create a hybrid token embedding for each word in the HPI note. The word embeddings, which are direct mappings from word x_i to vector \mathbf{t} , were pretrained using word2vec (Mikolov et al., 2013a;b;c) on all of the notes in MIMIC (v30) and only the discharge notes. Both the continuous bag of words (CBOW) and skip gram models were explored.

Let $z_{1:l}$ be the sequence of characters comprising the word x_i . Each character is mapped to its embedding \mathbf{c}_i , and all embeddings are input into a bidirectional LSTM, which ultimately outputs \mathbf{c} , the character embedding of

the word x_i .

The output of the token embedding layer is the vector \mathbf{e} , which is the result of concatenation of the word embedding, \mathbf{t} , and the character embedding, \mathbf{c} .

Contextual Representation Layer: The contextual representation layer takes as input the sequence of word embeddings, $\mathbf{e}_{1:k}$, and outputs an embedding of the contextual representation for each word in the HPI note. The word embeddings are fed into a bi-directional LSTM, which outputs h_j , a concatenation of the hidden states of the two LSTMs for each word.

Label Scoring Layer: At this point, each word x_i is associated with a hidden representation of the word, h_j . In the label scoring layer, we use a fully connected neural network with one hidden layer to output a score associated with each of the 10 categories for each word. Let $W \in R^{10 \times k}$ and $b \in R^{10}$. We can compute a vector of scores $\mathbf{s} = W \cdot h + b$ where the i th component of \mathbf{s} is the score of class i for a given word.

Label Sequence Optimization Layer: The Label Sequence Optimization Layer computes the probability of a labeling sequence and finds the sequence with the highest probability. In order to condition the label for each word on the labels of its neighbors, we employ a linear chain conditional random field (CRF) to define a global score, Q , for a sequence of words and their associated scores s_1, \dots, s_m and labels, y_1, \dots, y_n :

$$Q(y_1, \dots, y_n) = b[y_1] + \sum_{t=1}^m s_t[y_t] + \sum_{t=1}^{m-1} T[y_t, y_{t+1}] + \epsilon[y_m] \quad (1)$$

where T is a transition matrix $\in R^{10 \times 10}$ and $b, \epsilon \in R^{10}$ are vectors of scores that describe the cost of beginning or ending with a label.

The probability of a sequence of labels is calculated by applying a softmax layer to obtain a probability of a sequence of labels:

$$P(y_1, \dots, y_n) = \frac{e^{Q(y_1, \dots, y_n)}}{\sum_{y_1, \dots, y_n} e^{Q(y_1, \dots, y_n)}} \quad (2)$$

Cross-entropy loss, $-\log(P(\hat{y}))$, is used as the objective function where \hat{y} is the correct sequence of labels and the probability $P(\hat{y})$ is calculated according to the CRF.

3.3.3. RUNNING THE MODEL

We evaluated our model on the 515 annotated history of present illness notes, which were split in a 70% train set, 15% development set, and a 15% test set. The model is trained using the Adam algorithm for gradient-based optimization (Kingma & Ba, 2014) with an initial learning rate = 0.001 and decay = 0.9. A dropout rate of

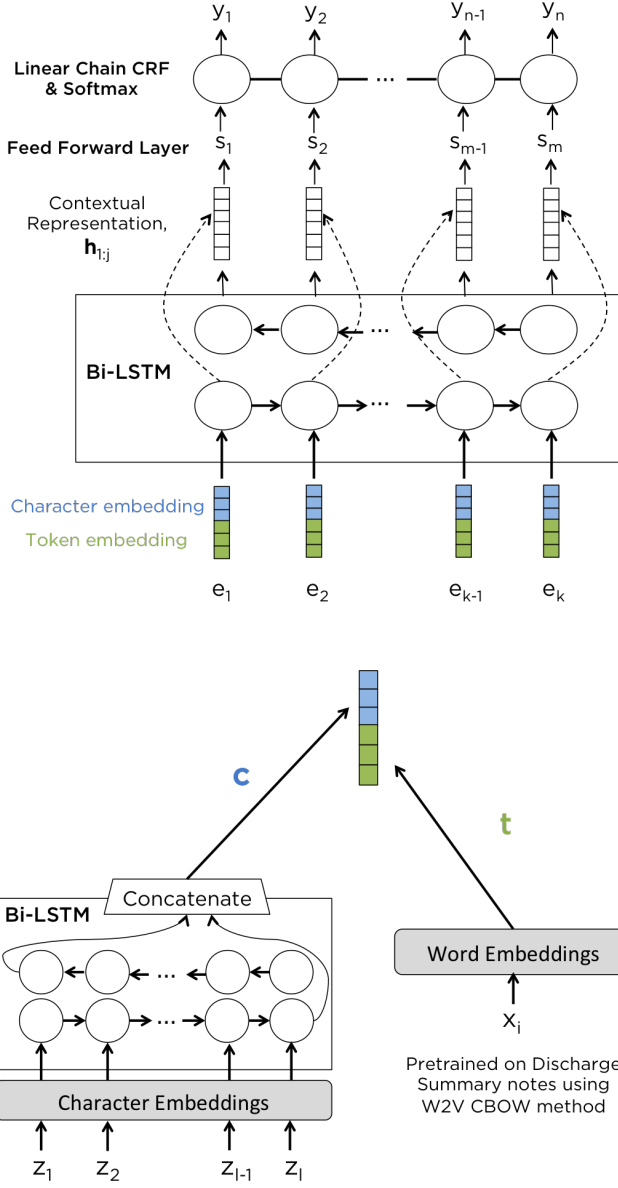


Figure 1. LSTM Model for sequential HPI word classification. X_i : token i , Z_1 to Z_l : characters of X_i , t : word embeddings, c : character embeddings, $e_{1:k}$: hybrid token embeddings, $h_{1:j}$: contextual representation embeddings, $s_{1:m}$: score vectors where $s[i]$ is the score for category i , $y_{1:m}$: output word labels.

0.5 was applied for regularization, and each batch size = 20. The model ran for 20 epochs and was halted early if there was no improvement after 3 epochs.

We evaluated the impact of character embeddings, the choice of pretrained w2v embeddings, and the addition of learned word embeddings on model performance on the dev set. We report performance of the best performing

model on the test set.

4. Results and Discussion

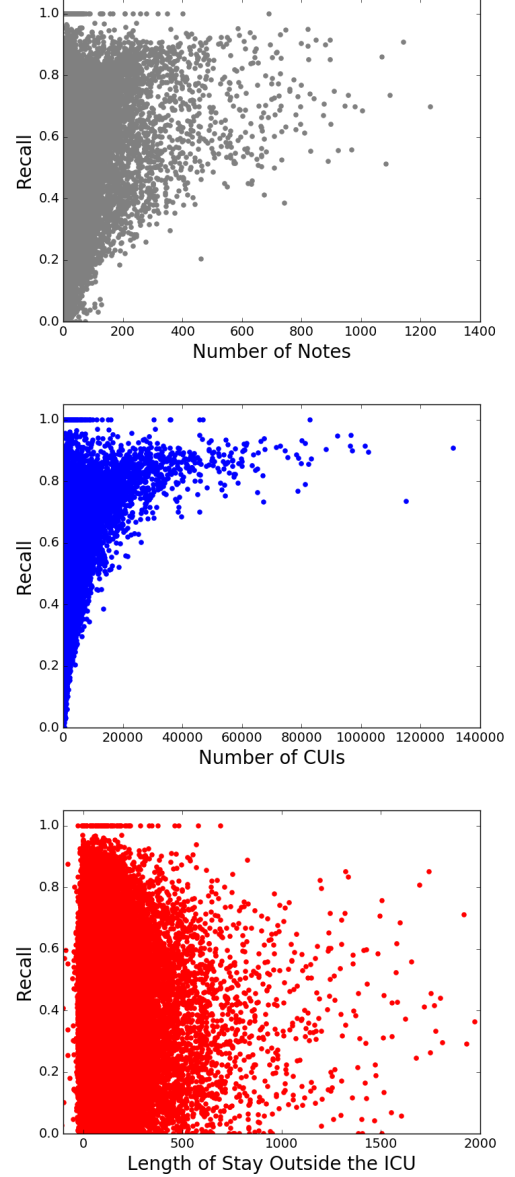


Figure 2. Relationship between the number of non-discharge notes (top), number of non-discharge CUIs (middle), or the number of hours the patient spent outside the ICU (bottom) and the recall for each discharge summary.

4.1. Upper Bound on Summarization

For each of the 55,177 discharge summary reports in the MIMIC database, we calculated what fraction of the CUIs in the discharge summary could be found in the remaining notes for the patient's current admission

(*by hadm_id*) and in their entire longitudinal medical record (*by subject_id*). Table 2 shows the CUI recall averaged across all discharge summaries by both *subject_id* and *hadm_id*. The low recall suggests that clinicians may incorporate information in the discharge note that had not been previously documented in the EHR. Figure 2 plots the relationship between the number of non-discharge notes for each patient and the CUI recall (top) and the number of total CUIs in non-discharge notes and the CUI recall (middle). The number of CUIs is a proxy for the length of the notes, and as expected, the CUI recall tends to be higher in patients with more and longer notes. The bottom panel in Figure 2 demonstrates that recall is not correlated with the patient’s length of stay outside the ICU, which indicates that our upper bound calculation is not severely impacted by only having access to the patient’s notes from their stay in the ICU.

Finally, Table 3 shows the recall for the sex, chief complaint, procedure, discharge medication, and HPI discharge summary sections averaged across all the discharge summaries. The procedure section has the highest recall of 0.807, which is understandable because procedures undergone during an inpatient stay are most likely to be documented in an EHR. The recall for each of these five sections is much higher than the overall recall in Table 2, suggesting that extractive summarization may be easier for some sections of the discharge note.

Overall, this upper bound analysis suggests that we may not be able to recreate a discharge summary with extractive summarization alone. While CUI comparison allows for comparing medically relevant concepts, cTAKES’s CUI labelling process is not perfect, and further work, perhaps through sophisticated regular expressions, is needed to define the limits of extractive summarization.

Table 2. Discharge CUI recall for each patient’s current hospital admission (*by hadm_id*) and their entire EHR (*by subject_id*), averaged across all discharge summaries

Comparison	Average Recall
By <i>subject_id</i>	0.431
By <i>hadm_id</i>	0.375

4.2. Labeling History of Present Illness Notes

Table 4 compares dev set performance of the model using various pretrained word embeddings, with and without character embeddings, and with pretrained versus learned word embeddings. The first row in each section is the performance of the model architecture described in the methods section for comparison. Models using word embeddings trained on the discharge summaries performed better than word embeddings trained

Table 3. Average Recall for five sections of the discharge summary. Recall for each patient’s sex was calculated by examining the structured data for the patient’s current admission, and recall for the remaining sections was calculated by comparing CUI overlap between the section and the remaining notes for the current admission.

Section	Average Recall
Sex	0.675
Chief Complaint	0.720
Procedure	0.807
Discharge Medication	0.580
HPI	0.665

on all MIMIC notes, likely because the discharge summary word embeddings better captured word use in discharge summaries alone. Interestingly, the continuous bag of words embeddings outperformed skip gram embeddings, which is surprising because the skip gram architecture typically works better for infrequent words (Google, 2013). As expected, inclusion of character embeddings increases performance by approximately 3%. The model with word embeddings learned in the model achieves the highest performance on the dev set (0.886), which may be because the pretrained word embeddings were trained on a previous version of MIMIC. As a result, some words in the discharge summaries, such as misspelled words or rarer diseases and medications, did not have associated word embeddings. Performing a simple spell correction on out of vocab words may improve performance with pretrained word embeddings.

We evaluated the best performing model on the test set. The Learned Word Embeddings model achieved an accuracy of 0.88 and an F1-Score of 0.876 on the test set. Table 5 shows the precision, recall, F1 score, and support for each of the ten labels, and Figure 3 shows the confusion matrix illustrating which labels were frequently misclassified. The demographics and patient movement labels achieved the highest F1 scores (0.96 and 0.93 respectively) while the vitals/labs and medication history labels had the lowest F1 scores (0.40 and 0.66 respectively). The demographics section consistently occurs at the beginning of the HPI note, and the patient movement section uses a limited vocab (transferred, admitted, etc.), which may explain their high F1 scores. On the other hand, the vitals/labs and medication history sections had the lowest support, which may explain why they were more challenging to label.

Words that belonged to the diagnosis history, patient movement, and procedure/results sections were frequently labeled as symptoms/signs (Figure 3). Diagnosis history sections may be labeled frequently as symptoms/signs because symptoms/diseases can either be de-

scribed as part of the patient’s diagnosis history or current symptoms depending on when the symptom/disease occurred. However, many of the misclassification errors may be due to inconsistency in manual labelling among annotators. For example, sentences describing both patient movement and patient symptoms (e.g. ”the patient was transferred to the hospital for his hypertension”) were labeled entirely as ’patient movement’ by some annotators while other annotators labeled the different clauses of the sentence separately as ’patient movement’ and ’symptoms/signs.’ Further standardization among annotators is needed to avoid these misclassifications. Future work is needed to obtain additional manual annotations where each HPI note is annotated by multiple annotators. This will allow for calculation of Cohen’s kappa, which measures inter-annotator agreement, and comparison of clinician and non-clinician annotator reliability.

Future work is also needed to better understand commonly mislabeled categories and explore alternative model architectures. Here we perform word level label prediction, which can result in phrases that contain multiple labels. For example, the phrase ”history of neck pain” can be labeled with both ’diagnosis history’ and ’symptoms/signs’ labels. Post-processing is needed to create a final label prediction for each phrase. While phrase level prediction may resolve these challenges, it is difficult to segment the HPI note into phrases for prediction, as a single phrase may truly contain multiple labels. Segmentation of sentences by punctuation, conjunctions, and prepositions may yield the best phrase chunker for discharge summary text.

Finally, supplementing the word embeddings in our LSTM model with CUIs may further improve performance. While word embeddings do well in learning the contextual context of words, CUIs allow for more explicit incorporation of medical domain expertise. By concatenating the CUI for each word with its hybrid token embedding, we may be able to leverage both data driven and ontology driven approaches.

5. Conclusion

In this paper we developed a CUI-based upper bound on extractive summarization of discharge summaries and presented a NN architecture that jointly classifies words in history of present illness notes. We demonstrate that our model can achieve excellent performance on a small dataset with known heterogeneity among annotators. This model can be applied to the 55,000 discharge summaries in MIMIC to create a dataset for evaluation of extractive summarization methods.

Table 4. F1-Scores on the development set under various model architectures. The first model in each section is the same model described in the methods section. Top performing models in each section are in bold.

Model	F1
w2v CBOW; Discharge Summaries	0.873
w2v Skip Gram; Discharge Summaries	0.831
w2v CBOW; All MIMIC Notes	0.862
w2v Skip Gram; All MIMIC Notes	0.809
With Character Embeddings	0.873
Without Character Embeddings	0.847
Pretrained Word Embeddings	0.873
Learned Word Embeddings	0.886

Table 5. Precision (P), Recall (R), F1 Score, and Number of supporting examples for each of the 10 label categories for the Learned Word Embedding model on the test set.

Label	P	R	F1	Support
Demographics	0.96	0.95	0.96	1489
DiagnosisHistory	0.83	0.82	0.83	3156
MedicationHistory	0.80	0.56	0.66	557
ProcedureHistory	0.85	0.87	0.86	1835
Symptoms/Signs	0.81	0.83	0.82	2235
Vitals/Labs	0.83	0.26	0.40	486
Procedures/Results	0.86	0.72	0.78	1114
Meds/Treatments	0.93	0.87	0.90	4165
PatientMovement	0.89	0.97	0.93	10933
Other	0.90	0.85	0.88	1815
Average/Total	0.88	0.88	0.88	27785

Confusion Matrix

Demographics	1421	28	0	0	2	6	1	0	16	0
DiagnosisHistory	51	2599	122	2	18	66	164	15	85	11
MedicationHistory	0	31	310	9	8	7	20	0	3	0
Meds_Treatments	0	14	47	1595	21	15	36	58	29	57
Movement_Admit_Transf	0	27	9	61	1858	53	25	133	113	21
Other	0	0	15	3	1	127	0	0	7	0
ProcedureHistory	1	44	13	10	12	2	802	42	3	6
Procedures_Results	0	14	0	38	54	23	26	3626	59	73
Symptoms_Signs	16	388	38	73	239	185	33	266	10567	99
Vitals_Labs	0	11	3	44	22	2	7	25	51	1548

Predicted

Demographics
DiagnosisHistory
MedicationHistory
Meds_Treatments
Movement_Admit_Transf
Other
ProcedureHistory
Procedures_Results
Symptoms_Signs
Vitals_Labs

True

Figure 3. Confusion Matrix for the Learned Word Embedding model on the test set. True labels are on the x-axis, and predicted labels are on the y-axis.

Acknowledgments

We would like to thank our annotators, Andrew Goldberg, Laurie Alsentzer, Elaine Goldberg, Andy Alsentzer, Grace Lo, and Josh Donis. We would also like to acknowledge Pete Szolovits for his guidance and for providing the pretrained word embeddings and Tristan Naumann for providing the MIMIC CUIs.

References

- Allahyari, Mehdi and Kochut, Krys. Semantic tagging using topic models exploiting wikipedia category network. In *Semantic Computing (ICSC), 2016 IEEE Tenth International Conference on*, pp. 63–70. IEEE, 2016.
- Allahyari, Mehdi, Pouriyeh, Seyedamin, Assefi, Mehdi, Safaei, Saeid, Trippe, Elizabeth D, Gutierrez, Juan B, and Kochut, Krys. Text summarization techniques: A brief survey. *arXiv preprint arXiv:1707.02268*, 2017.
- Alsentzer, L. personal communication.
- Christensen, T. and Grimsno, A. (n.d.). Instant availability of patient records, but diminished availability of patient information: A multi-method study of gps use of electronic patient records. *BMC Medical Informatics and Decision Making*, 8(12), 2008.
- Dernoncourt, Franck, Young Lee, Ji, and Szolovits, Peter. Neural networks for joint sentence classification in medical paper abstracts. pp. 694700, 12 2016.
- Edmundson, H. P. New methods in automatic extracting. *J. ACM*, 16(2):264–285, April 1969. ISSN 0004-5411. doi: 10.1145/321510.321519. URL <http://doi.acm.org/10.1145/321510.321519>.
- Feblowitz, J. C., Wright, A., Singh, H., Samal, L., and Sittig, D. F. Summarization of clinical information: A conceptual model. *Journal of Biomedical Informatics*, 44:688–699, 2011.
- Ferreira, Rafael, de Souza Cabral, Luciano, Lins, Rafael Dueire, e Silva, Gabriel Pereira, Freitas, Fred, Cavalcanti, George DC, Lima, Rinaldo, Simske, Steven J, and Favaro, Luciano. Assessing sentence scoring techniques for extractive text summarization. *Expert systems with applications*, 40(14):5755–5764, 2013.
- Goldstein, A. and Shahar, Y. An automated knowledge-based textual summarization system for longitudinal, multivariate clinical data. *Journal of Biomedical Informatics*, 61(Supplement C):159 – 175, 2016. ISSN 1532-0464. doi: <https://doi.org/10.1016/j.jbi>.
- 2016.03.022. URL <http://www.sciencedirect.com/science/article/pii/S1532046416300156>.
- Google. Google Code Archive word2vec, 2013. URL <https://code.google.com/archive/p/word2vec/>.
- Kingma, Diederik and Ba, Jimmy. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.
- Kripalani, Sunil, Lefevre, Frank, Phillips, Christopher, V Williams, Mark, Basaviah, Preetha, and Baker, David. Deficits in communication and information transfer between hospital-based and primary care physicians. *JAMA : the journal of the American Medical Association*, 297:831–41, 03 2007.
- Liu, P. and Pan, X. Google Research Blog text summarization with tensorflow, 2016. URL <https://research.googleblog.com/2016/08/text-summarization-with-tensorflow.html>.
- Luhn, H. P. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2(2):159–165, April 1958. ISSN 0018-8646. doi: 10.1147/rd.22.0159. URL <http://dx.doi.org/10.1147/rd.22.0159>.
- Mikolov, Tomas, Chen, Kai, Corrado, Greg, and Dean, Jeffrey. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013a.
- Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg S, and Dean, Jeff. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pp. 3111–3119, 2013b.
- Mikolov, Tomas, Yih, Wen-tau, and Zweig, Geoffrey. Linguistic regularities in continuous space word representations. In *hlt-Naacl*, volume 13, pp. 746–751, 2013c.
- Moen, H., Peltonen, L.M., Heimonen, J., Airola, A., Pahikkala, T., Salakoski, T., and Salanterä, T. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67(Supplement C):25 – 37, 2016. ISSN 0933-3657. doi: <https://doi.org/10.1016/j.artmed>. 2016.01.003. URL <http://www.sciencedirect.com/science/article/pii/S0933365716000051>.
- Nallapati, Ramesh, Zhou, Bowen, Gulcehre, Caglar, Xiang, Bing, et al. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*, 2016.

- Natarajan, K., Stein, D., Jain, S., and Elhadad, N. (n.d.). An analysis of clinical queries in an electronic health record search utility. *International Journal of Medical Informatics*, 79(7):515–522, 2010.
- NLM. *UMLS Reference Manual*. National Library of Medicine, 2009.
- Paulus, Romain, Xiong, Caiming, and Socher, Richard. A deep reinforced model for abstractive summarization. *arXiv preprint arXiv:1705.04304*, 2017.
- Pivovarov, R. *Electronic Health Record Summarization Over Heterogenous and Irregularly Sampled Clinical Data*. PhD thesis, Department of Biomedical Informatics, Columbia University, 2016.
- Rim, Kyeongmin. Mae2: Portable annotation tool for general natural language use. in proceedings of the 12th joint acl-iso workshop on interoperable semantic annotation. 05 2016.
- Rush, Alexander M, Chopra, Sumit, and Weston, Jason. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015.
- Savova, Guergana K, Masanz, James J, Ogren, Philip V, Zheng, Jiaping, Sohn, Sunghwan, Kipper-Schuler, Karin C, and Chute, Christopher G. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.
- Singh, H., Spitzmueller, C., Petersen, N. J., Sawhney, M. K., and Sittig, D. F. (n.d.). Information overload and missed test results in electronic health record-based settings. *JAMA Internal Medicine*, 173(8), 2013.