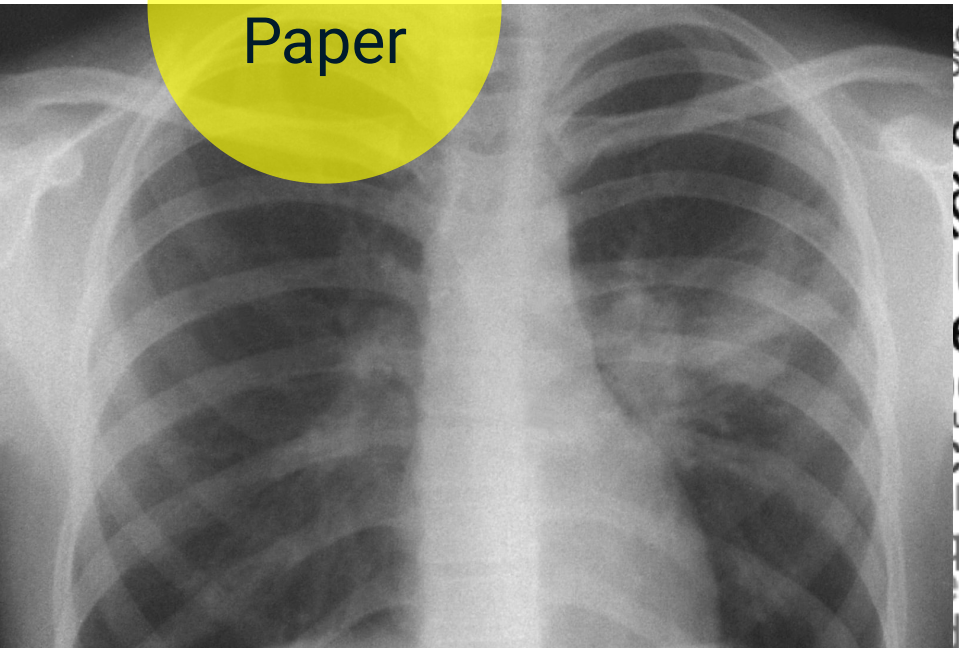
COMPUTATIONAL LINGUISTICS:

A NOVEL APPROACH TO IDENTIFY PULMONARY NODULES AND EXTRACT THEIR CHARACTERISTICS FROM RADIOLOGY REPORTS

Natalia Rodnova, Lead Data Scientist, Eon
August 2020

White Paper



ABSTRACT

We developed a Computational Linguistics (CL) information extraction (IE) model that identifies the dimension, location, and characteristics of the largest lung nodule mentioned in a free-text radiology report. The model is based on a set of rich linguistic rules and an in-house knowledge base. It was developed on ~20,000 randomly selected radiology reports from 18 institutions. The manually created gold standard consisted of 2,480 reports with 2,070 true positive and 410 true negative results. The overall accuracy between the model and the gold standard for the presence of a lung nodule was 98.95%, with 98.99% precision and 99.66% recall. The measurement accuracy of the largest nodule was 97.74%. In this document, we provide an introduction to CL and IE modeling in particular; briefly describe related publications; outline our modeling approach; and review pertinent results.

COMPUTATIONAL LINGUISTICS

> "Human knowledge is expressed in language. So computational linguistics is very important."

Mark Steedman, ACL Presidential Address [1]

Computational Linguistics (CL) is a relatively new interdisciplinary field that creates computer systems capable of understanding, analyzing, and extracting meaning from written and spoken language. It is based on traditional Linguistics, Statistics, Computer Science (CS), and Machine Learning (ML). CL, in conjunction with knowledge representation and formal reasoning theories, creates a foundation for Artificial Intelligence (AI).

Once a topic of science fiction, the ability of machines to use human language is now fundamental to many applications that are integral parts of our daily lives. Web search engines, email spam filters, translation software, dictation modules on our phones and computers, digital voice assistants, and chatbots are examples of CL applications.

CL studies several, somewhat overlapping, domains. They include machine translation (MT), automatic speech recognition (ASR), text-to-speech (TTS), information extraction (IE), natural language understanding (NLU), natural language generation (NLG), conversational NLP, and ontology learning (OL).

While each subfield of CL has its applications in healthcare, we focus our paper on Information Extraction, one of the most actively developing areas. IE turns the unstructured information embedded in text into structured data, for example, for populating a relational database or creating a time series [2]. The model presented in this paper is an IE model and we will cover its components in the **EON Model Methods** section.

IE BUILDING BLOCKS

A typical IE system performs the following steps [2]: tokenization and lemmatization, sentence boundary detection, part-of-speech tagging/dependency parsing, named entity recognition/disambiguation, coreference resolution, relation extraction, temporality normalization, and template filling. Most of these tasks are intuitive and easy for humans to perform, but for computer systems, are quite challenging. Below is a brief description of each task.

Tokenization: Breaking text into a sequence of tokens representing words, punctuation, acronyms, URLs, etc. This is the most basic task, and multiple "out-of-the-box" solutions exist that perform tokenization well. Some adjustments to out-of-the-box solutions are needed to adapt them to the clinical domain and to take into consideration specifics of the language used in EHRs.

Lemmatization: Finding the base form of words, or words that can be looked up in a dictionary. Some challenges that exist include determining which part of speech a word belongs to: "I am a doctor (am => to be)" OR "It is 5 am (am => a.m.)"

Sentence Boundary Detection: A well-implemented task for general text, such as literature, periodicals, web articles, and social media posts. However, in the clinical domain, there are challenges due to nonstandard or ambiguous abbreviations, incomplete sentences, use of jargon, lack of punctuation caused by either omission or by optical character recognition (OCR) software, and more.

Part-of-Speech Tagging (POS): Determining the part of speech each token in a sentence is (noun, verb, adjective, adverb, etc.).

Dependency Parsing: Determining syntactic dependencies between words in a sentence (subject, object, adjectival modifier, etc.). POS tagging and dependency parsing are well-developed on general text. However, these models perform rather poorly on clinical narratives in part because of specific vocabulary, irregular writing style, and a lack of annotated data sets for training models.

Named Entity Recognition (NER): A subtask of IE, NER aims to locate and classify named entities mentioned in unstructured text into predefined categories such as person names, organizations, locations, medical codes, time expressions, quantities, etc.

Named Entity Disambiguation: A task of assigning a unique identity to entities mentioned in text. That identity associates an input text fragment, identified as a named entity, to a corresponding unique entity in a target knowledge base. For example, an "apple" can be the fruit produced by an apple tree (Malus domestica) and be identified by Germplasm Resources Information Network (GRIN) Taxonomy #104681 [6]. Alternatively, it can also refer to Apple, Inc. – a corporation registered with the Security and Exchange Commission (SEC) and identified by SEC CIK #0000320193 [7].

Coreference Resolution: The task of determining whether two mentions in the text co-refer, i.e. refer to the same entity in the discourse model (said another way, the same discourse entity) [2]. For example: "There is a ground-glass opacity in the left lung. Today it measures 7 mm. On the previous study, the area measured 5 mm". In the 2nd and 3rd sentences, "it" and "the area" both refer to "a ground-glass opacity" mentioned in the 1st sentence.
Relation Extraction: The task of extracting relationships between identified named entities (located in, part of, has characteristic, etc.)

Temporality Detection and Normalization: Determining when the events described in text occurred and aligning the relative expressions like "on the previous study" or "last year" with the context baseline time.

Template Filling: Organizing extracted data in a structured form.

There are two additional fundamental models that are often used in conjunction with IE: word frequency models and the word embeddings model. We used both in exploratory data analysis as well as for developing our linguistic rules—these are covered in more detail on the following pages.

WORD FREQUENCY MODELS

The two most basic tools used in CL for information extraction are bag-of-words (BOW) and TF-IDF (Term Frequency - Inverse Document Frequency) models.

Bag-of-words is based on calculating word occurrences in text documents. Figure 1 shows a visual representation (called a word cloud) of the most frequent 500 words in 20K chest CT reports. The size of the words is proportional to their frequency in the documents.

Figure 1. Word cloud built on word frequency from 20K chest CT reports. The image was generated using https://www.wordclouds.com/

On the other hand, TF-IDF evaluates how important a word is to a document. Its value increases proportionally to the number of times the word appears in the document and is offset by the number of documents containing the word.

WORD EMBEDDINGS MODEL

Word embeddings model builds a multidimensional representation of the words based on their semantic similarity [8]. It is an unsupervised model (requiring no human involvement for data annotation or feature engineering) but requires a lot of contextual data. We trained our word embeddings model on 8M radiology reports. Figure 2 contains an intuitive visualization of the word embedding model. Nine seed words were selected (lung, nodule, carcinoma, ground glass, pneumonia, image, patient, implant, firefighter), and for each of them, the 25 most similar words were identified by the model. After reducing the dimensionality from 150 to 2, the word embeddings were plotted (Figure 2). The coordinates of the dots on this graph reflect the word's semantic roles. On the right side of the graph, the "lung," "ground glass," and "pneumonia" clusters (purple, teal, green) are distinct but are not completely separated from each other. Also, while the "nodule" cluster (lavender) is mostly on the upper left, close to the "carcinoma" cluster (blue), there are very few words (nodules and nodularities) that are on the right-hand side of the graph and pretty close to the "lung" cluster.

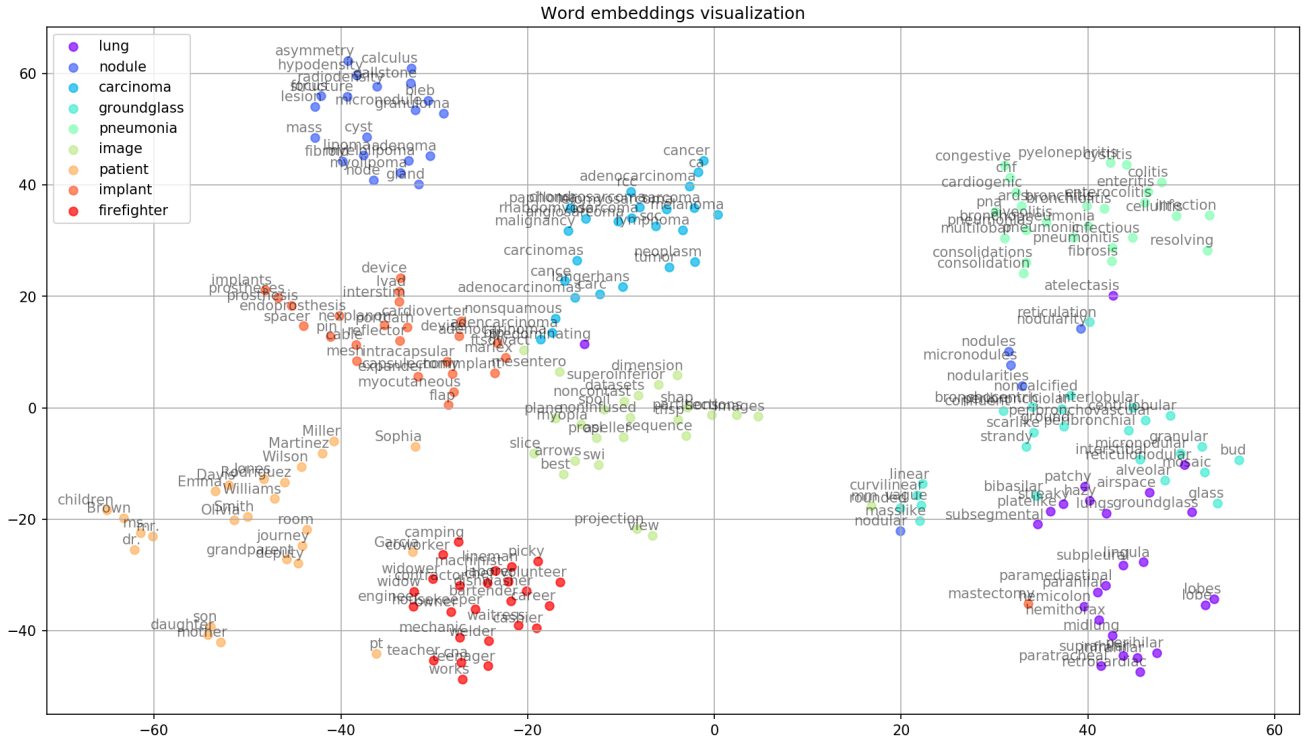


Figure 2. Word embeddings visualization. (Note: first and last names in "patient" cluster were replaced by the most frequent names in the US [5])

Word embeddings (vectors) are used to calculate the semantic similarity between words and phrases. They are also used as input features to most of the IE sub-models described above.

RELATED WORK

There are a number of research publications studying clinical IE systems and several review articles. In one review [3], 263 publications from 2009 to 2016 were identified, including 43 papers with radiology reports as the data source. Another review [4] reported 63 systems analyzing radiology reports.

We reviewed the literature and identified the most recent, relevant publications with the highest performance to use for benchmarking. The following four were the best match to the objectives of our model, and are presented in order from the least similar to most similar to our model:

1 The earliest relevant work was by Kaiser Permanente, published in 2013 [9]. They created a basic, rule-based model to identify radiology reports containing mention of lung nodules. The model combined diagnostic codes, procedure codes, and text mining. The model was run on about 7,000 reports of health plan members who underwent chest CT exams after receiving a diagnosis code indicating the possible presence of lung nodules. The model did not extract nodule measurements or characteristics, it was a binary classification model (nodule present or absent). Based on ~100 annotated reports, the model's precision was 87% and recall 96%

2 More similar to ours, a model to extract and categorize finding measurements was created in 2015 by Philips (Healthcare) Research [10]. This model extracted organ measurements using regular expressions and categorized them into three groups—a measurement of a clinical finding, relative position (distance), and technical spec (image, slice, etc.). Then measurement temporality was determined using an ML method. The model did not determine which clinical finding the specific measurement was describing nor the location of the finding. The model was evaluated on 2,000 sentences (not documents), all from a single institution. Measurement extraction precision was 99.4% and recall was 99.1%. Accuracy of the classification (location of the finding within the organ) was 96%.

3 In 2016, a NER model was developed at Stanford University [11] that extracted five categories from radiology reports—organs, organ locations, clinical findings, characteristics, and expressions of uncertainty. This model performance was the lowest among these four publications, with 87.7% precision and 82.9% recall, averaged over the 4 classifiers.

4 In 2019, another model was developed by Stanford researchers [12]—this is the most similar to our model. Their objective was to extract measurements and their temporality, organ, and characteristics. The model was based on a combination of rules and ML, similar to ours. It was developed on 1K records from a single institution, with 100 records manually annotated. The measurement-only accuracy was at 97%, however the model performance deteriorated quickly with the addition of other modifiers: measurement + organ: 83%; measurement + organ + abnormality: 80%; and measurement + organ + abnormality + characteristic: 66%.

EON MODEL METHODS

Our goal was to create a model for extraction of the measurement, location, and characteristics of the largest lung nodule mentioned in a radiology report. To develop this model, we selected a dataset of 20,000 radiology reports mentioning lung findings from 18 different institutions. Based on this dataset, we developed a collection of linguistic rules using various CL techniques, ranging from bag of words (BOW) to word embeddings.

We used a combination of a neural-network-based ML model for standard CL tasks, including POS-tagging and dependency parsing, and rules-based linguistics (for the NER component) to create our model.

Our rules are not regular expressions (something found in most natural language understanding models), but are linguistic rules combining words, their forms and characteristics (part of speech, plural/singular form, verb tense, etc.) and dependencies (subject, object, modifier, etc.), as well as their location within the document.

To accomplish this, we used a proprietary ontology to assign meanings to words and determine their relations. The ontology is based on fragments of the Foundational Model of Anatomy (FMA) [13], RadLex [14], and was enriched with the knowledge of our subject matter experts (SMEs). The visualization of a small subset of the ontology is shown in Figure 3.

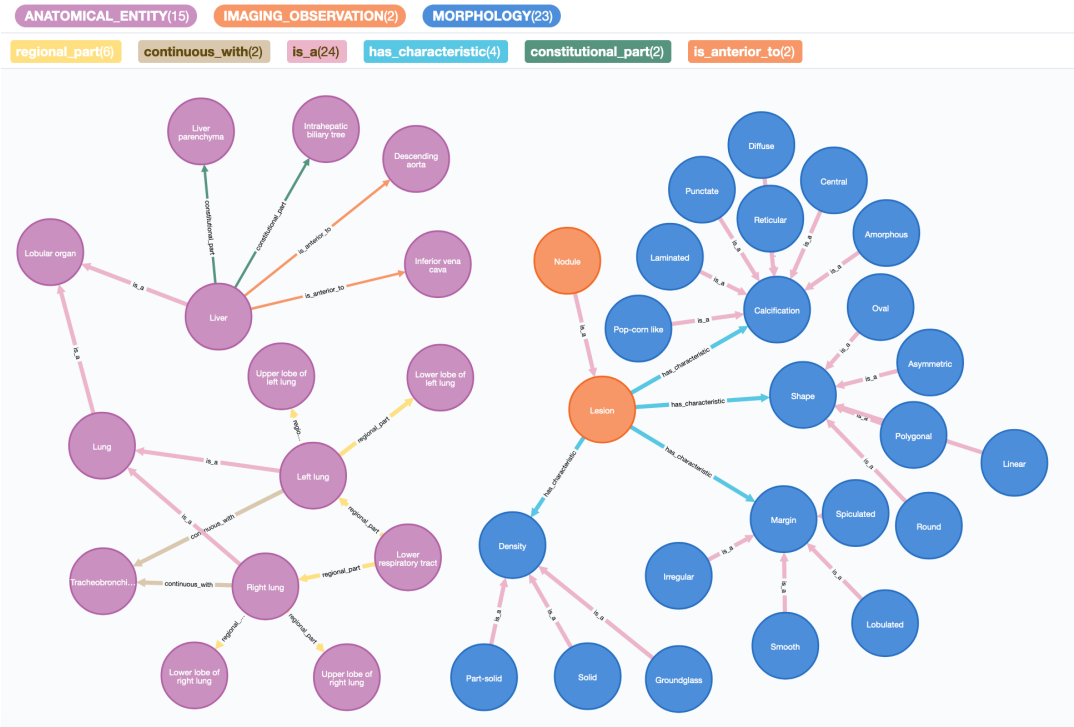


Figure 3. Knowledge base (ontology) representation graph.

The most complicated step of the model was defining the relationships between entities. After all previous steps are completed, and the model knows what the tagged entities (measures, organs, locations) are referring to, as well as knows the syntactic dependencies between them, the model uses a set of rules and heuristics to determine semantic relations between entities.

Once the relations extraction is completed, the model has all the required information to extract data of interest, in this case, a set of measured lung nodules. It excludes measurements from reference citations (such as the Fleischner Society [15] guidelines for incidentally detected lung nodule follow-up). Finally, the model selects the largest nodule and outputs its size, location, and characteristics.

MODEL EVALUATION

To evaluate the model performance, we selected 2,480 radiology reports from CT exams that covered at least part of the lungs. Our two SMEs, an interventional pulmonologist and an imaging physicist, annotated these documents in parallel, with all annotation conflicts reviewed and resolved afterward. The annotations were performed on maximum dimension size of the largest lung nodule. The dataset contains 2,070 true positive documents (with lung nodules present) and 410 true negative documents (without lung nodules).

PARAMETER	VALUE
Max measurement	8 mm
Single/multiple	Multiple
Laterality	Bilateral
Max Nodule Lobe	Lower
Max Nodule Shape	Round
Max Nodule Density	Solid
Max Nodule Margin	Smooth
Max Nodule Calcification	Non-calcified

Table 1. Final model output with sample values.

Smaller subsets of this dataset were annotated for the characteristics of the largest nodule as shown in Table 1.

We found the accuracy agreement between the model and the annotated gold standard records for the presence of a lung nodule was 98.95%, with 98.99% precision and 99.66% recall. The measurement accuracy of the measurement of the largest nodule was 97.74%. The accuracy results did not decrease with the additional classification of nodule characteristics; conversely, accuracy was found to improve slightly.

REFERENCES:

[1] Mark Steedman. 2008. Last Words: On becoming a discipline. Computational Linguistics, Vol. 34, #1.

[2] Dan Jurafsky and James H. Martin. 2019. Speech and Language Processing. Edition 3, unpublished. https://web.stanford.edu/~jurafsky/slp3/

[3] Yanshan Wang et al. 2018. Clinical information extraction applications: A literature review. Journal of Biomedical Informatics, Vol 77, Jan 2018, pp 34-49.

[4] Ewoud Pons, et al. 2016. Natural Language Processing in Radiology: A Systematic Review. Radiology. 2016 May; 279(2):329-43. https://pubs.rsna.org/doi/10.1148/radiol.16142770

[5] List of most common surnames in North America. Wikipedia. wikipedia.org.

[6] U.S. National Plant Germplasm System. npgsweb.ars-grin.gov

[7] Security and Exchange Commission Filing SEC CIK #0000320193. https://sec.report/CIK/0000320193

[8] Tomas Mikolov, et al. 2013. Efficient Estimation of Word Representations in Vector Space. https://arxiv.org/abs/1301.3781

[9] Kim N. Danforth. 2013. Automated Identification of Patients with Pulmonary Nodules in an Integrated Health System Using Administrative Health Plan Data, Radiology Reports, and Natural Language Processing. J Thorac Oncol. 2012 Aug; 7(8): 1257–1262. doi: https://www.jto.org/article/S1556-0864(15)32691-5/fulltext

[10] M. Sevenster. 2015. Natural Language Processing Techniques for Extracting and Categorizing Finding Measurements in Narrative Radiology Reports. Applied Clinical Informatics 2015; 6: 600–610. doi: https://www.thieme-connect.de/products/ejournals/abstract/10.4338/ACI-2014-11-RA-0110

[11] Saeed Hassanpoura and Curtis P. Langlotz. Information extraction from multi-institutional radiology reports. Artif Intell Med. 2016 January; 66: 29–39. doi: https://www.sciencedirect.com/science/article/pii/S0933365715001244?via%3Dihub

[12] Selen Bozkurt, et al. 2019. Automated Detection of Measurements and Their Descriptors in Radiology Reports Using a Hybrid Natural Language Processing Algorithm. Journal of Digital Imaging (2019) 32:544–553. doi: https://link.springer.com/article/10.1007%2Fs10278-019-00237-9

[13] Rosse C., Mejino J.L.V. 2008. The Foundational Model of Anatomy Ontology. In: Burger A., Davidson D., Baldock R. (eds) Anatomy Ontologies for Bioinformatics. Computational Biology, vol 6. Springer, London. doi: https://link.springer.com/chapter/10.1007/978-1-84628-885-2_4

[14] Curtis P. Langlotz, et al. 2006. RadLex: A New Method for Indexing Online Educational Materials. RadioGraphics Vol. 26, No. 6. doi: https://pubs.rsna.org/doi/10.1148/rg.266065168

[15] H. MacMahon, et al. Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. Radiology 2017; 284(1): 228-243.