Identifying Semantics in Clinical Reports Using Neural Machine Translation

Srikanth Mujjiga, Vamsi Krishna, Kalyan Chakravarthi, Vijayananda J
HealthSuite Insights, Philips Healthcare, Bangalore, India
{srikanth.mujjiga,vamsi.krishna,kalyan.chakravarthi.murahari,vijayananda.j}@philips.com

Abstract

Clinical documents are vital resources for radiologists when they have to consult or refer while studying similar cases. In large healthcare facilities where millions of reports are generated, searching for relevant documents is quite challenging. With abundant interchangeable words in clinical domain, understanding the semantics of the words in the clinical documents is vital to improve the search results. This paper details an end to end semantic search application to address the large scale information retrieval problem of clinical reports. The paper specifically focuses on the challenge of identifying semantics in the clinical reports to facilitate search at semantic level. The semantic search works by mapping the documents into the concept space and the search is performed in the concept space. A unique approach of framing the concept mapping problem as a language translation problem is proposed in this paper. The concept mapper is modelled using the Neural machine translation model (NMT) based on encoder-decoder with attention architecture. The regular expression based concept mapper takes approximately 3 seconds to extract UMLS concepts from a single document, where as the trained NMT does the same in approximately 30 milliseconds. NMT based model further enables incorporation of negation detection to identify whether a concept is negated or not, facilitating search for negated queries.

Motivation

The adoption of Health Information Systems (Winter et al. 2011), Picture Archiving and Communication Systems and Electronic Medical Records by healthcare facilities has resulted in large amount of clinical documents in digital form. The availability of large amount of clinical documents in digital format is transforming the healthcare facilities from volume based care providers to value based care providers. The push is significantly more in the radiology department, where the radiologists are increasingly using patient's clinical history and consult the past similar cases for clinical comparisons and past outcomes to provide targeted healthcare (McEnery 2018). For example, a radiologist might be interested in consulting the reports of past male patients above 40 years of age, having no smoking habit and diagnosed with throat cancer. Likewise a clinical researcher might be interested in knowing the number of findings of benign tumour in patients who have undergone computed tomography scans.

Information retrieval (IR) is the de facto way to leverage the knowledge available in large resources of clinical documents like radiology reports, discharge summaries, electronic medical records, etc. However, the effectiveness of the IR system has a significant impact on the clinical and academic productivity of the radiologists and clinical researchers. Building effective IR systems for clinical text is challenging because of large number of clinical terms with varied use. With abundant interchangeable words in clinical domain, understanding the semantics of the words in the clinical documents is vital to improve the search results. For example, search for "heart attack" should be able to retrieve the documents containing "myocardial infarction" as they both mean the same. The lexicon based IR systems are futile in clinical domain due to different ways of expressing the same clinical findings. The IR system that retrieves and prioritizes the clinical reports based on the semantics of the reports is more likely to provide better search results, rather than the one based on lexicons. For example, "cyst found in left lower lung" is an observation similar to "calcification of thorax" and effective IR system should be able to identify that both the observations are similar.

Related Work

There is an extensive amount of research to identify clinical entities and the relationship between these entities (Raja, Subramani, and Natarajan 2013), (Rak et al. 2012). Authors (Wei 2017) proposed a mechanism to map gene variants to unique identifiers. These text mining approaches form the basis on which other applications can be build, to make use of the underlying knowledge. One such application developed by (Kurnit et al. 2017) is a search engine specialised in cancer therapies. However, these works are focussed on addressing the information retrieval problem of the clinical subdomains. They is very sparse work around productisable IR system in clinical domain which can cater to multiple modalities and clinical literature.

(Batet, Sánchez, and Valls 2011) and (Mabotuwana, Lee, and Cohen-Solal 2013) have computed the document similarity of radiology reports by using knowledge-based semantic similarity using ontologies.

Due to the computational limitations in processing large

size ontologies, most of the research to calculate semantic similarity using clinical knowledge base is restricted to single or few ontologies. The knowledge based IR system's response time is proportional to the number of ontologies used. Most of the knowledge based IR system's make use of only few ontologies and have very limited applications as they do not generalize outside those ontologies.

Clinical Semantic Search

In clinical domain many phrases or words semantically produce the same meaning but are syntactically different. For example, the phrases heart attack, cardiovascular disease, congestive heart failure and myocardial infarction all mean same. Over the years researchers in healthcare domain have developed large number of ontologies like International Classification of Diseases (ICD), Medical Subject Headings (MeSH), RadLex, SNOMED CT etc.

Unified Medical Language System (UMLS) developed by US National Library of Medicine, brings together many health and biomedical vocabularies and standards to enable interoperability (Bodenreider 2004). It is a comprehensive thesaurus of concepts spreading over health information, medical terms, drug names, and billing codes. The UMLS contains around 11 million distinct concepts with over 3.6 million names from more than 200 contributing sources in 25 languages. It also includes 12 million relations (Schulze-Kremer, Smith, and Kumar 2004) among these concepts making it richest knowledge base in clinical domain.

The clinical semantic search explained in this paper works by mapping the documents into the UMLS concept space. The search is performed in the concept space for document retrieval. Figure 1 shows the high level modules in the clinical semantic search. As with any IR system, it involves the indexing step to index the clinical documents and the search step for document retrieval. The pseudocode for indexing workflow is shown in algorithm 1 and the search workflow is shown in algorithm 2.

```
Data: Clinical Text
Result: Index the input Clinical Text
all_concepts ← ∅;
sections ← Section_Detector(Clinical Text);
for sentence in sections do
    concepts ← Concept_Mapper(sentence);
    negated_entities ← Negx_Detector(sentence);
    for concept in concepts do
        if concept['entity'] in negated_entities then
            concept['cui'] ← concept['cui'] + "_1";
        else
            concept['cui'] ← concept['cui'] + "_0";
        end
    end
    all_concepts = all_concepts + concepts;
end
Index all_concepts into the Search Engine;
```

Algorithm 1: Indexing Workflow

```
Data: Search Query
Result: Relavent Documents
concepts ← Concept_Mapper(Search Query);
negated_entities ← Negx_Detector(Search Query);
for concept in concepts do
    if concept['entity'] in negated_entities then
        concept['cui'] ← concept['cui'] + "_1";
    else
        concept['cui'] ← concept['cui'] + "_0";
    end
end
Search concept in the Search Engine Index;
```

Algorithm 2: Search Workflow

Challenge

The most challenging work in developing clinical semantic search engine is building the concept mapper which maps the clinical sentences to enriched UMLS concepts. UMLS along with the ontologies includes a number of tools, one among them is Metamap (Aronson 2001) which can be used to extract UMLS concepts from the clinical text. MetaMap uses a knowledge-intensive approach based on symbolic, natural-language processing (NLP) and computational-linguistic techniques. With the full UMLS (all ontologies) loaded, MetaMap takes approximately 3 seconds to extract concepts from a single document. This makes use of Metamap impractical for productionisation.

The challenge is to develop a concept mapper which can be an alternative for MetaMap with acceptable accuracy and performance. Table 1 shows examples of clinical sentences and few of the corresponding concepts. The matched words in the table are the words due to which the concept was assigned to the corresponding sentence. The goal is to develop a concept mapper which given the clinical sentence as input will generate the corresponding concepts.

Concept mapping using NMT

The problem of machine translation deals with conversion of text from one language to other with no human intervention. Machine translation task using neural models received major attention due to recent research breakthroughs in deep neural network architectures (Bahdanau, Cho, and Bengio 2014). Neural based machine translation has outperformed benchmarks which were mostly by traditional machine learning and rule based algorithms (Vaswani et al. 2017).

The proposed technique frames the concept mapping problem as a machine translation problem where the source language is the clinical text in english and the target language is the enriched UMLS concepts. The concepts are enriched because they include the negation information denoting whether the concepts are negated or not. The NMT model is trained to translate clinical sentences in english to enriched UMLS concepts. The dotted block in figure 1 is replaced with NMT based code mapper.

A simple NMT system consists of two components:

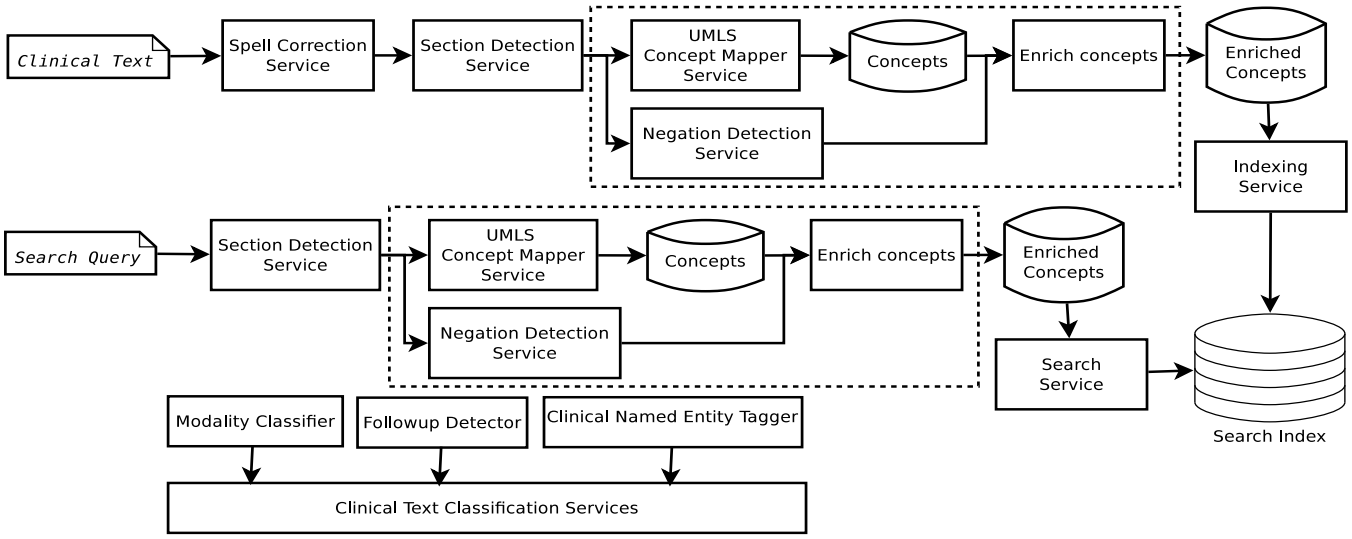

Figure 1: High Level Modules

Clinical Sentences	Concepts	Matched Words
Tools that help doctors see inside the blood vessels may help prevent heart attack	C0027051 C0018787 C0005847	heart attack heart blood vessels
Ventricular septal rupture is a rare but feared complication after myocardial infarction	C0242875 C0027051	ventricular septal rupture myocardial infarction
The conservative treatment of kidney cancer in middle aged and elderly patients	C0007134 C0022646 C0006826	kidney cancer kidney cancer
Suramin is not an active agent in advanced renal carcinoma	C0038880 C0007134 C0007097	suramin renal carcinoma carcinoma

Table 1: Clinical sentences and few of their corresponding concepts along with the matched words

- An encoder component that computes the representation of source sentence into a fixed hidden vector.
- A decoder component that generates target representation.

Recurrent Neural Networks (RNNs) (Rumelhart, Hinton, and Williams 1988) and Convolution Neural Networks (CNNs) (Lecun et al. 1989) can be used to model encoders. A decoder is modelled using RNNs. Attention mechanism (Luong, Pham, and Manning 2015) is the recent area of research in machine translation achieving state of the art results. Attention improves the NMT by focusing on parts of the source sentence while decoding the target sentence. The attention weights determine which part of the source sentence to focus on during translation. These weights are learned using a neural network. Different NMT architectures empirically evaluated for mapping clinical text to enriched concepts are listed below.

1. RNN as both encoder and decoder.
2. CNN as encoder and RNN as decoder.
3. RNN and CNN as encoder and decoder modelled using RNN.
4. RNN as both encoder and decoder along with attention layer (general attention).
5. RNN and CNN as encoder and RNN as decoder along with tweaks to attention layer.

Data Preparation

Training an NMT model to map clinical sentences to UMLS concepts needs large number of annotated sentences, annotations being the tagged enriched UMLS concepts. Table 2 shows two sample annotated sentences used for building the NMT based concept mapper.

The Text Retrieval Conference (TREC) runs Clinical Decision Support (CDS) track every year focusing on the retrieval of clinical documents such as clinical reports, published biomedical articles and medical records (Roberts et al. 2016). Every year as part of TREC-CDS track, large number of clinical documents are released. These documents are abstracts, largely from Medline/PubMed (730k to 1.25 million text articles), clinical trial studies provided by U.S. National Library of Medicine and de-identified clinical notes of actual patients. 5 million sentences are extracted from 9 GB

Input Sentences	Output Concepts			
fracture dislocation of the ankle with the fibula fixed behind the tibia.	0C0016068 0C0016658 0C3714578	0C0040184 0C1281580	0C0159877 0C0434691	0C0012691 0C1279118
infants placed on the waterbed during the first four postnatal days benefited more than those placed later.	0C0021270 0C0814225 0C0205087 0C0205450	0C0442504 0C0443281 0C0205172	0C1704765 0C0439228 0C0205435	0C1882509 0C0438858 0C1279901

Table 2: Clinical sentences and few of their corresponding concepts along with the matched words

compressed text files released by TREC-CDS - 2018.

Metamap along with our negation detection model is used to annotate 5 million sentences. The negation detection model enhances the concepts with the negation information. The data annotator took 7 days to generate the annotations. All the concepts which occur less than 50 times are ignored. This left us with 1.5 millions enriched concepts. Table 2 shows two such generated annotated sentences. 70-20-10 split is used for training, validation and testing.

Leveraging Clinical Embeddings

Word embeddings are the distributed vector representation of the words (Bengio et al. 2003). Word embeddings have proven to provide state of the art performances in many natural language processing tasks (Lample et al. 2016). Pre trained word embeddings released by google contain 300-dimensional vectors for 3 million words. However, these embeddings are not well suited for clinical domain as they lack many clinical words as well as the semantic similarity with respect to clinical domain. TREC-CDS data is used to do transfer learning on the skip-gram model (Mikolov et al. 2013) to fine tune the google embeddings to clinical context.

The Transfer learning helped significantly to capture clinical context. Figure 2 shows few clusters of the word embeddings after transfer learning. The clusters are formed using euclidean distance as similarity measure. The green borders drawn manually in the figure 2 shows the clusters that are semantically close in the clinical domain. These embeddings are used to initialise the embedding layer of NMT at the encoder stage (for both RNN and CNN).

Attention Mechanism

The modifications are made to the general attention of (Luong, Pham, and Manning 2015). Figure 3 shows the attention based architecture using both CNN and RNN as the encoder. The idea behind having CNN encoder in addition to RNN encoder is to model n-gram information. CNN encoder will encode the phrase level information in the sentence and this information is used in the attention layer along with the RNN encoder at every time step during the decoding to get the appropriate attention weights. Attention layer along with input feeding approach is used at every time step of decoder. The highlighted block in Figure 3 details the attention mechanism. The alignment vector a_t whose size equals the number of time steps of the encoder is derived by comparing the current target hidden state h_t, hidden vector from CNN

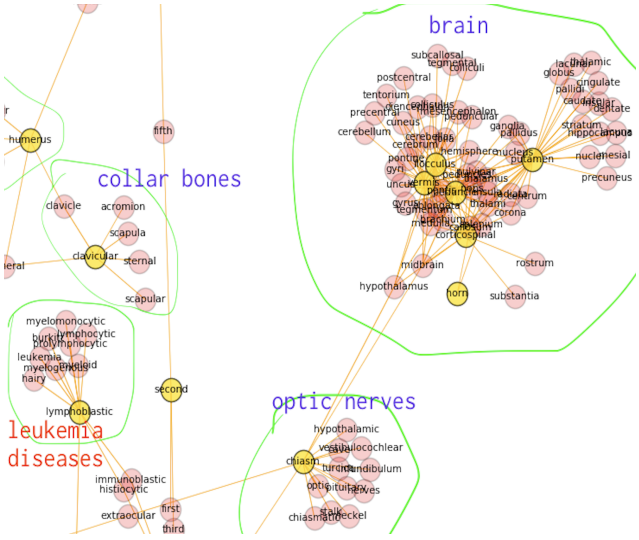

Figure 2: Clinical Word Embeddings (word2vec)

h_{conv} and the hidden vectors of all the encoder time steps represented by $\hat{h}_s$.

$$a_t(s) = \frac{exp(score(h_t, h_{conv}, \hat{h}_s))}{\sum_{s^i} exp(score(h_t, h_{conv}, \hat{h}_{s^i}))} \quad where \qquad (1)$$

$$score(h_t, h_{conv}, \hat{h}_s) = (h_t, h_{conv})^T W_a \hat{h}_s \qquad (2)$$

At each time step t of decoder the model computes the alignment vector based on the current target state, all the source states of RNN encoder and CNN encoder state. The context vector c_t is the weighted average of the alignments weights which are computed using equation 1 with the encoder hidden vectors $\hat{h}_s$. This context vector along with the current hidden target state of decoder is given to the generator to generate the output.

Evaluation

Table 3 lists the different architectures evaluated. Column Ppl in the table represents the perplexity and BLEU represents the Bilingual Evaluation Understudy score (Papineni et al. 2002)on the test data.

The RNN encoders are modelled using an LSTM encoder of sequence length 50 with 4 stacked layers. For CNN encoder, 500 filters with various filter dimensions (4X300,

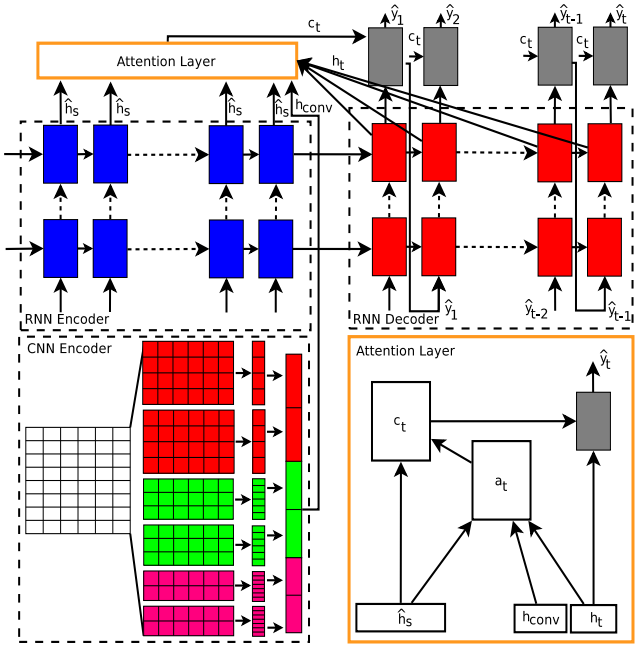

Figure 3: NMT with Attention

3X300, 2X300) are used. The hidden vector sizes of both RNN and CNN encoders are of size 500. The RNN decoder is modelled using LSTM with 4 stacked layers. Dropout layer with probability 0.3 and adam with learning rate of 0.001 is used. The β_1 and β_2 parameters of adam are set to 0.9 and 0.999 respectively. The models are implemented in PyTorch (Paszke et al. 2017) and trained on a single Nvidia TITAN X GPU. The best performing model in table 3 was trained with batch size of 64 for 7 days to converge.

The overall performance of semantic search is measured using the Mean Average Precision (MAP) metric @10. 5000 clinical reports across five modalities namely Diagnostic X-Ray, mammography, magnetic resonance (MR) imaging, ultrasound and computer tomography are collected. 50 relevant queries across all modalities are created with the help of three radiologists. With the help of clinical experts, 10 ranked relevant reports for each query are selected from the 5000 reports. The 5000 reports are indexed into the semantic search and the MAP is calculated for the top 50 queries with the top 10 search results. With NMT model base concept mapper, clinical semantic search achieved MAP@10 score of 0.76.

Figures 4 shows the home screen of semantic search indexed with the MR radiology reports. Figures 5 shows the rendering page of the de-identifed clinical report. As shown in the figure, the clinical Named Entity Recognition (NER) models and clinical document classifiers are used to enhance the user experience. For example, an anatomy tagger is used to highlight all the anatomies in the rendered report.

Conclusion and Future Work

This paper presented an approach to build clinical IR system based on the semantics. This paper emphasis the importance of the method to map clinical sentences to concepts and how it effects the functional and non functional aspects of the underlying system. We presented an approach to model this concept mapping challenge as a machine translation problem. We evaluated various NMT model architectures for concept mapping and observed that the CNN+RNN as encoder and RNN as decoder with modified attention layer gave us the best results. However, the CNN+RNN as encoder model took more time in training when compared to RNN only encoder models.

Improvements in the NMT model require reindexing of the data. An efficient incremental indexing pipeline can solve this problem to certain extent. Also from our experience, we observed that the users are more interested in user based customisations to the search results. To address this problems we are working on building a learning system based on the users feedback.

We observed that the NMT model fails to capture negated concepts when the negation cue and the focus word are far apart. The analysis showed that the percentage of negated concepts is very low in the training dataset. As future work we plan to address this problem by data augmentation and model enhancements. As of now we have used 1.5 million concepts as target vocabulary, we plan to extend the vocabulary to cover more concepts.

Figure 4: Home screen

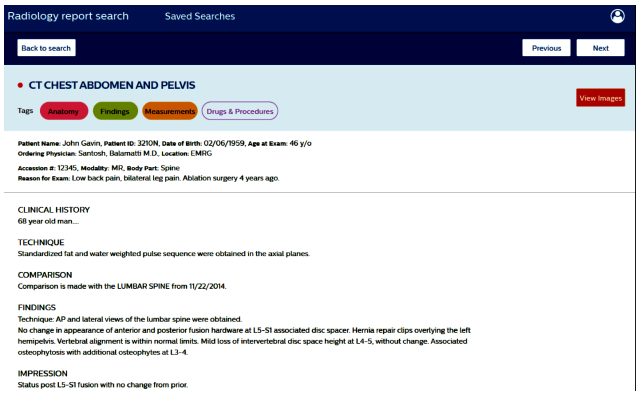

Figure 5: Report Rendering Page

No	Model Architecture	Ppl	BLEU
1	RNN Encoder & RNN Decoder	5.33	15.1
2	RNN Encoder & RNN Decoder + dropout	5.01	18.1
3	RNN Encoder & RNN Decoder + dropout + input feeding	5.01	18.5
4	RNN Encoder & RNN Decoder + dropout + input feeding + clinical embeddings	4.97	20.5
5	CNN Encoder & RNN Decoder	5.96	14.0
6	CNN Encoder & RNN Decoder + dropout	5.05	17.5
7	CNN Encoder & RNN Decoder + dropout + input feeding	5.04	17.5
8	CNN Encoder & RNN Decoder + dropout + input feeding + clinical embeddings	4.98	19.0
9	**RNN Encoder & RNN Decoder + dropout + input feeding + clinical embeddings + attention**	**3.15**	**38.5**
10	CNN Encoder + RNN Encoder & RNN Decoder	5.12	16.4
11	CNN Encoder + RNN Encoder & RNN Decoder + dropout	5.00	18.8
12	CNN Encoder + RNN Encoder & RNN Decoder + dropout + input feeding	4.92	20.0
13	CNN Encoder + RNN Encoder & RNN Decoder + dropout + input feeding + clinical embeddings	4.67	22.3
14	**CNN Encoder + RNN Encoder & RNN Decoder + dropout + input feeding + clinical embeddings + attention**	**3.03**	**40.1**

Table 3: Evaluation of different model architectures

References

Aronson, A. R. 2001. Effective mapping of biomedical text to the UMLS metathesaurus: The MetaMap program. *Proceedings of the AMIA Symposium* 17–21.

Bahdanau, D.; Cho, K.; and Bengio, Y. 2014. Neural machine translation by jointly learning to align and translate. *CoRR* abs/1409.0473.

Batet, M.; Sánchez, D.; and Valls, A. 2011. An ontology-based measure to compute semantic similarity in biomedicine. *J. of Biomedical Informatics* 44(1):118–125.

Bengio, Y.; Ducharme, R.; Vincent, P.; and Janvin, C. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.* 3:1137–1155.

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 32:D267–D270.

Kurnit, K.; Bailey, A.; Zeng, J.; Johnson, A.; Shufean, M.; Brusco, L.; Litzenburger, B.; Sanchez, N.; Khotskaya, Y.; Holla, V.; Simpson, A.; Mills, G.; Mendelsohn, J.; Bernstam, E.; Shaw, K.; and Meric-Bernstam, F. 2017. Personalized cancer therapy: A publicly available precision oncology resource. *Cancer Research* 77(21):e123–e126.

Lample, G.; Ballesteros, M.; Subramanian, S.; Kawakami, K.; and Dyer, C. 2016. Neural architectures for named entity recognition. *CoRR* abs/1603.01360.

Lecun, Y.; Jackel, L.; Boser, B.; Denker, J.; Graf, H.; Guyon, I.; Henderson, D.; Howard, R.; and Hubbard, W. 1989. Handwritten digit recognition: Applications of neural network chips and automatic learning. 27:41 – 46.

Luong, M.; Pham, H.; and Manning, C. D. 2015. Effective approaches to attention-based neural machine translation. *CoRR* abs/1508.04025.

Mabotuwana, T.; Lee, M. C.; and Cohen-Solal, E. V. 2013. An ontology-based similarity measure for biomedical data - application to radiology reports. *J. of Biomedical Informatics* 46(5):857–868.

McEnery, K. W. 2018. *Reference Guide in Information Technology for the Practicing Radiologist*. American College of Radiology.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318. Stroudsburg, PA, USA: Association for Computational Linguistics.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in pytorch.

Raja, K.; Subramani, S.; and Natarajan, J. 2013. Ppinterfinder–a mining tool for extracting causal relations on human proteins from literature. *Database : the journal of biological databases and curation* 2013:bas052.

Rak, R.; Rowley, A.; Black, W.; and Ananiadou, S. 2012. Argo: an integrative, interactive, text mining-based workbench supporting curation. *Database : the journal of biological databases and curation* 2012:bas010.

Roberts, K.; Demner-Fushman, D.; Voorhees, E. M.; and Hersh, W. R. 2016. Overview of the trec 2016 clinical decision support track. *Proceedings of Text Retrieval Conference 2016*.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1988. Neurocomputing: Foundations of research. Cambridge, MA, USA: MIT Press. chapter Learning Representations by Back-propagating Errors, 696–699.

Schulze-Kremer, S.; Smith, B.; and Kumar, A. 2004. Revising the umls semantic network. In *MedInfo*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. *CoRR* abs/1706.03762.

Wei, C.-H. 2017. Tmvar 2.0: Integrating genomic variant information from literature with dbsnp and clinvar for precision medicine. 34.

Winter, A.; Haux, R.; Ammenwerth, E.; Brigl, B.; Hellrung, N.; and Jahn, F. 2011. *Strategic Information Management in Hospitals*. London: Springer London. 237–282.