THE DESIGN OF AN ONCOLOGY KNOWLEDGE BASE FROM AN ONLINE HEALTH FORUM

by

Omar Ramadan

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Electrical and Computer Engineering

Indianapolis, Indiana

May 2022

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Zina Ben Miled, Chair

Department of Electrical & Computer Engineering

Dr. Paul Salama

Department of Electrical & Computer Engineering

Dr. Euzeli Cipriano Dos Santos

Department of Electrical & Computer Engineering

Approved by:

Dr. Brian King

To Mom and Dad,

ACKNOWLEDGMENTS

This research was supported in part by Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., Kenilworth, NJ, USA. I would like to thank Weilin Meng, Dr. Christopher M. Black, and Dr. Lixia Yao from Merck Co., Inc. for their guidance and valuable input during this project. I would also like to thank Jarod Baker of the Regenstrief Institute for his support. Finally, my gratitude and appreciation go to Dr. Zina Ben Miled, my thesis advisor.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

KB	knowledge base
OKB	oncology knowledge base
NTN	neural tensor network
IRN	implicit reasoNet
RNN	recurrent neural network
NER	named entity recognition
BERT	bidirectional encoder representations from transformers

ABSTRACT

Knowledge base completion is an important task that allows scientists to reason over knowledge bases and discover new facts. In this thesis, a patient-centric knowledge base is designed and constructed using medical entities and relations extracted from the health forum r/cancer. The knowledge base stores information in binary relation triplets. It is enhanced with an *is-a* relation that is able to represent the hierarchical relationship between different medical entities. An enhanced Neural Tensor Network that utilizes the frequency of occurrence of relation triplets in the dataset is then developed to infer new facts from the enhanced knowledge base. The results show that when the enhanced inference model uses the enhanced knowledge base, a higher accuracy (73.2 %) and recall@10 (35.4%) are obtained. In addition, this thesis describes a methodology for knowledge base and associated inference model design that can be applied to other chronic diseases.

1. INTRODUCTION

Knowledge bases (KB) have been used in a number of applications to enable data-driven deductions and rule-based inferences. KBs can contain general information such as real-world facts (e.g., Freebase [1]) or they can be domain-specific (e.g., Hepatitis Knowledge Base [2]). In general, KBs are developed to support decision-making applications including recommender [3] and question answering systems [4]. KBs can also be used for knowledge discovery as in the case of this study. Common to all these applications is the need for a methodology than can augment the KB. This is necessary because KBs are often incomplete [5]. Knowledge base completion is a technique that can enrich the content of the KB by automatically inferring new facts using existing ones [5].

In this work, an Oncology Knowledge Base (OKB) is constructed from the Reddit r/-cancer online health forum [6]. Subscribers to r/cancer often post detailed information related to their experience with cancer. This information may include treatment choices, treatment side-effects, comorbid conditions, and symptoms. An advantage of the Reddit forums over other social media networks is that they are anonymous [7]. This anonymity affords patients and caregivers the ability to openly discuss their journey with the disease and how it affected them physically or mentally.

The aim of this study is to extract information submitted by the users relating to their experience with the disease from a subset of the r/cancer posts in the form of relation triplets. Posts often detail the user's experience in an anecdotal form. Consequently, oncology-relevant information must be extracted using natural language processing methods such as Named Entity Recognition (NER) to create the relation triplets. Once this information is collected, an enhanced KB is constructed to represent the extracted data. This KB becomes the source of data that is used for training and testing a machine learning model capable of inferring new facts. The addition of a hierarchical relation between semantically and/or syntactically related entities (e.g., lung cancer and cancer) to the OKB is investigated in order to compare the inference ability of the KB completion model with and without this type of relation.

The inference model is based on the Neural Tensor Network (NTN) architecture [8]. Because the subreddit r/cancer contains posts addressing similar topics, duplicate relation triplets were found across several posts. In order to account for duplicate triplets, the inference model was trained on a version of the OKB that tracks the frequency of each relation triplet. The performance of this enhanced NTN model is compared to that of the baseline model where the frequency of each triplet is ignored and only occurrence matters.

The main contributions of the present thesis include:

- The construction of an Oncology Knowledge Base that consists of relevant medical entities and the relations between those entities as identified from posts on the health forum r/cancer;
- The development of a KB completion model that can help answer oncology-related questions from the perspective of the patient or the caregiver;
- A comparative analysis of the performance of an enhanced NTN model to that of the baseline NTN on two versions of the OKB: one that contains hierarchical relations and one that does not.

2. RELATED WORK

KBs are a collection of entity-relation pairs where entities represent objects in the target domain and relations represent links between the objects. Binary entity-relation models store information in the form <head entity, relation, tail entity> where each relation semantically connects an entity pair, a head and a tail, to form a triplet (e.g., <Chemotherapy, is-a-treatment-of, Cancer>). The entities and the relations are encoded using a numerical vector and the space of all encodings is referred to as the embedding space. KBs usually contain a large number of triplets. However, they may still be incomplete [5]. Several previous studies suggested different models for KB augmentation or completion by inferring new triplets from existing ones [8]–[14].

One of the main goals of this study is to develop and augment a KB from an online health forum, specifically r/cancer. Health-related KBs and ontologies are available. For example, UMLS [15] is an ontology for biomedical concepts. Similarly, OMIM [16] compiles information on human genetics. These KBs are created from sources such as biomedical literature and clinical notes. However, there is no oncology-specific KB that is derived from an online health forum. KB completion models can be generally classified under three categories: additive, multiplicative, and neural-network-based.

2.1 Additive Models

TransE [10] is a type of additive model that learns the translations from the head entity to the tail entity [17]. The scoring function in TransE for a given triplet <h,r,t> is as follows:

$$S(h, r, t) = \|h + r - t\|, \tag{2.1}$$

where h and t are the embeddings of the head and tail entities, and r is the embedding of the relation [10]. If the triplet <h,r,t> exists in the KB, then the model will attempt to minimize the score for the triplet during training [11]. One problem that this model faces is its lack of ability to express reflexive, one-to-many, many-to-one, and many-to-many relations [11]. TransE is an example of a translation-based model. Several variants of TransE have

been proposed in the literature. In particular, the variants aim to modify the translations in order to increase the inference abilities of the model for different relation types. One variant, TransH [11], allows for a different expression of entity embeddings when present in different relations. This is done by creating a relation-specific hyper-plane for each relation in the embedding space, and projecting the embeddings of the head and tail entities onto that hyper-plane. The translation vector that connects the head and tail projections on the hyper-plane is the representation of the relation between them. The scoring function of TransH for a given triplet <h, r, t> is as follows:

$$f_r(h,t) = \left\| h - w_r^T h\, w_r \right) + d_r - (t - w_r^T t\, w_r) \right\|_2^2, \tag{2.2}$$

where f_r is the relation-specific scoring function, h and t are the embeddings of the head and tail entities, w_r is the normal vector to the relation hyper-plane, and d_r is the relation translation between the head and tail entities on the relation hyper-plane. TransE and TransH share a common approach in that the entity and relation embeddings share the same embedding space. Another variant is TransR [14] which creates different embedding spaces, one for entities and R embedding spaces for relations, where R denotes the number of relations in the KB. The head and tail entity embeddings are projected from the entity space to the relation space using a mapping function as follows:

$$h_r = h\, M_r, \quad t_r = t\, M_r, \tag{2.3}$$

where h and t represent the embeddings of the head and tail entites, h_r and t_r represent the projections of the head and tail embeddings onto the embedding space of relation r, and M_r represents the projection matrix. After the projection is complete, a scoring function identical to that of TransE (Equation 2.1) is used to compute the likelihood of the triplet <h, r, t> being true. One problem that translational models face is that they do not utilize the structure of the KB to reason over transitive triplets [14]. Specifically, this problem becomes prominent when dealing with health-related data where various symptoms, diseases, treatments, and side-effects are related across multiple relations.

2.2 Multiplicative Models

DistMult [9] is a multiplicative bilinear model that represents relations as matrices in the embedded vector space which interact multiplicatively with the entity embeddings [17]. It is a simplified variant of NTN with limited expressive power as it can only model linear interactions [8]. The scoring function which is optimized during the development of the KB completion model is defined as:

$$S(h, R, t) = h^T W_R t \,, \tag{2.4}$$

where h and t represent the vector embeddings for the head and tail entities, and W_R is a relation-specific matrix. It was shown in [9] that the operational choice of using a multiplicative approach by DistMult outperformed the additive one by TransE on benchmark datasets. In DistMult, W_R is implemented as a diagonal matrix. That is, the order in which the head and tail entities appear in the relation does not matter. This forces all relations to be symmetric which limits the ability of the model to express uni-directional relations. ComplEx [12] is another example of a KB completion model that belongs to the multiplicative model category. It extends DistMult by embedding both entities and relations in the complex embedding space [18]. The entity embeddings interact with the relation embeddings using the Hermetian dot product which involves the conjugate-transpose of one of the two vectors. This approach allows for distinct representations of relations where the head and tail are interchanged.

2.3 Neural-Network-Based Models

Implicit ReasoNets (IRNs) [13] is a KB completion model that falls under the neural-network-based model category. During the training process, the model learns the tail entity embedding given the head entity and the relation (<h, r, ?>). The model utilizes an encoder to create head entity and relation embeddings which are concatenated to form an intermediate representation. A module called the controller forms the main component of the model. It consists of a recurrent neural network (RNN), an attention mechanism, and a

decision-making module. The RNN employs an iterative process whereby the intermediate representation is computed at each step. During this iterative process, a shared memory is used to store and retrieve relevant information about the KB using the attention mechanism. At every step subsequent to the first one, the controller uses the output of the RNN and the shared memory to generate the next intermediate representation. The controller decides when to stop creating intermediate representations and produce a final prediction vector for the tail entity using a probability measure dependent on the intermediate state.

Since most relations in the medical field are many-to-many (e.g., is a symptom of), it was important to select a model capable of reasoning over many-to-many relations. Several of the models mentioned earlier (e.g., [9], [12], [13]) have this capability. However, NTN [8], a neural-network-based KB completion model, was selected for the augmentation of the proposed OKB because of its expressive power. It utilizes a combination of linear and bilinear operators in addition to the non-linearity in its scoring function [9]. Each relation has its own weight matrix containing multiple slices. Using multiple slices allows for the encoding of many-to-many relations. The scoring function used in NTN (Equation 2.5), relates the embeddings of the entities and those of the relations by using tensor multiplication [17]. A bilinear layer is used to connect the head and tail entities along with the relation between them across multiple dimensions. Another interaction between the head and tail entities is present in the linear layer where their embeddings are concatenated. NTN encodes information about entities on the word level. The entity embeddings are formed by averaging the embeddings of their constituent words. For example, the embedding of the entity *Brain Cancer* is formed using the average of the embeddings of the words *Brain* and *Cancer*. The scoring function in NTN for a given triplet <h, r, t> is as follows:

$$S(h, R, t) = u_R^T f\left(h^T W_R^{[1:k]} t + V_R \begin{bmatrix} h \\ t \end{bmatrix} + B_R\right), \tag{2.5}$$

where f is the hyperbolic tangent activation function; h and t are the embeddings of the head and tail entities, $W_R^{[1:k]}$ is a tensor with k slices where each slice represents a different instantiation of the relation R, V_R and u_R are weight matrices, and B_R is a bias term [8].

NTN was competitive on some benchmarks compared to bilinear multiplicative models when the comparison is based on accuracy [8]. However, NTN underperformed when compared to other models such as IRNs using other benchmarks and the evaluation metric mean rank [13].

3. METHODOLOGY

The proposed framework consists of two main components: The OKB and the KB completion model. This chapter describes the data preprocessing steps needed to construct the OKB followed by the methodology used to develop and validate the KB completion model.

3.1 Oncology Knowledge Base

Binary relations are extracted from r/cancer posts in the form of triplets consisting of <head, relation, tail>. The head and tail entities are medical entities. They vary according to the content of the post. However, the list of relations is limited to a set of relations that are of interest to the current study, namely, *is-a-treatment-of, is-a-symptom-of, is-a-side-effect-of, co-existing-symptoms, and co-existing-diagnosis*. The first three relations are uni-directional relations while the last two are bi-directional.

The NER function of Comprehend Medical [19] was used to identify the medical entities from the posts. Comprehend Medical is a collection of several machine learning models that are trained on clinical text notes [19]. Specifically, the NER functionality of Comprehend Medical is able to identify and organize medical concepts from free text into categories [20]. Three of these categories are used in this study and they are defined in Comprehend Medical as follows:

- Category 1: "Signs, symptoms, and diagnosis of medical conditions."
- Category 2: "Methods that are used to determine a medical condition."
- Category 3: "Medication and dosage-related information."

Entities from the above categories are further stratified into subcategories called traits, types, and attributes. These sub-categories provide more specific information about the entity. For instance, if an entity belongs to the first category, the traits subcategory can be used to determine whether the entity is a sign, symptom, or diagnosis. Similarly, if an entity belongs to the second category, the types subcategory can be used to determine if the entity is a test, treatment, or procedure. Using the combination of categories and subcategories

provided by Comprehend Medical, the proposed OKB is organized according to the following set of entity identifiers and their definitions [20]:

- Sign: "A medical condition that the physician reported."
- Symptom: "A medical condition reported by the patient."
- Diagnosis: "A medical condition that is determined as the cause or result of the symptoms."
- Treatment: "Interventions performed over a span of time for combating a disease or disorder. This includes groupings of medications, such as antivirals and vaccinations."
- Procedure: "Interventions as a one-time action performed on the patient to treat a medical condition or to provide patient care."
- Medication: "Medication and dosage information for the patient."

Comprehend Medical was selected for NER because it provides an automated method for extracting and categorizing medical entities. Moreover, it was trained on clinical notes making it suitable for the current study. Alternate NER systems include BioALBERT [21] and ScispaCy [22]. That said, the fact that Comprehend Medical was trained on clinical notes rather than health forum posts can still be a limitation for the current study. Medical notes are usually expressed over a formal medical lexicon whereas posts are informal and often use a casual writing style.

Once entities are identified, the second step in the OKB development is the creation of triplets. To achieve this, a blueprint was established for each relation. These blueprints are expressed in terms of constraints based on the category of the head and tail entities in a given relation as shown below:

- < Medication/Treatment/Procedure, is-a-treatment-of, Diagnosis >
- < Symptom/Sign, is-a-symptom-of, Diagnosis >
- < Symptom/Sign, is-a-side-effect-of, Medication/Treatment/Procedure >

- < Symptom/Sign, co-existing-symptoms, Symptom/Sign >
- < Diagnosis, co-existing-diagnosis, Diagnosis >

For example, for the relation *is-a-symptom-of*, the head entity can only belong to the categories Symptom or Sign while the tail entity can only belong to the Diagnosis category. The purpose of the blueprints is to generate meaningful triplets from the context of each post. Once NER is performed on the text of the post, the resulting entities are matched to the applicable blueprint in order to generate all possible triplets associated with a given post. For example, if one entity belongs to the Symptom category and another to the Treatment category within the post, a triplet is created with the first entity as the head, the second entity as the tail, and *is-a-side-effect-of* as the relation (e.g., <hair loss, is-a-side-effect-of, chemotherapy>. A given post can be associated with a large number of triplets. A randomly selected subset of 1,000 posts from all the r/cancer posts [6] in 2020 were selected for this study and the OKB was developed from the content of these posts.

The final step in the data processing consists of the manual annotation of the triplets to develop the initial knowledge base. Three annotators were presented with each post and corresponding relation triplets. Using only the content of the post and no additional knowledge, they were then asked to classify the triplets into one of three classes:

- Relation exists: The relation can be inferred from the context of the post.
- Relation does not exist: The relation cannot be inferred from the context of the post.
- Suspected relation: The patient/physician/caretaker mentions that the relation is possible with no explicit confirmation or assertation.

The majority vote of the 3 annotators is taken as the final class for each relation. A subject matter expert reviewed the annotations for additional quality control. The annotation guidelines presented to the annotators for this task are shown in Appendix A.

The triplets that belong to the *relation exists* class are used to build the OKB. The remaining triplets are discarded. Moreover, only unique triplets are included in the OKB. Multiple occurrences of the same triplet are used to assign a weight to the triplet. Table 3.1

shows example posts with the corresponding entities, relation triplets and triplet annotation. In the first post, < IV antibiotics, is-a-treatment-of, nocardia infection> is a valid triplet where *IV antibiotics* and *nocardia infection* are entities that were identified by Comprehend Medical. The former belongs to the treatment category while the latter belongs to the diagnosis category. These categories match the blueprint of the *is-a-treatment-of* relation. Therefore, a triplet was formed using these two entities and the *is-a-treatment-of* relation. The annotators read the post and asserted that the relation is inferred from the post as indicated by the Triplet Class in Table 3.1.

The two triplets < IV Antibiotics, is-a-treatment-of, Nocardia infection > and < Nocardia infection, co-existing-diagnosis, Abscess > in post 1 are annotated as relation exists. According to the post, the patient suffered from a nocardia infection from their Hickman line. The Hickman line was then removed, and the patient was put on a course of IV Antibiotics. It is clear that the patient was administered IV antibiotics to treat their nocardia infection; hence, the first relation is true. The nocardia infection also caused an abscess. This gives us enough evidence to rule that the second relation is also true. As for the third relation < Chemo, is-a-treatment-of, Nocardia infection >, there's no evidence in the post to suggest that the patient used chemo to treat their nocardia infection; hence, the relation cannot be confirmed.

In post 2, the patient was diagnosed with cancer and subsequently experienced several symptoms such as throwing up and falling unconscious. From the context of the post, it is clear that throwing up was a symptom of cancer, hence, the triplet < Throwing up, is-a-symptom-of, Cancer> is a true relation. Additionally, it is mentioned that the patient was observed to experience the symptoms throwing up and falling unconscious around the same time frame. Thus, the triplet < Unconscious, co-existing-symptoms, Throwing up > is a also a true relation.

In post 3, the patient was diagnosed with Adenocarcinoma and it is suspected that the patient may also suffer from Lynch Syndrome. Thus, the triplet < Adenocarcinoma, co-existing-diagnosis, Lynch Syndrome > is classified under suspected relation.

Biomedical entities extracted from a social media forum can suffer from variations that can limit the potential of any KB completion method. In order to address this limitation, a new *is-a* relation is introduced. This relation is used for three types of variations:

- Syntactic variations such as misspelled (e.g., Hodkin's Lymphoma for Hodgkin's Lymphoma), pluralized (e.g., pains for pain), and shortened (e.g., chemo for chemotherapy) forms of the entities.
- Semantic variations where different entities refer to the same semantic concept (e.g., "lose weight" and "weight loss").
- Hierarchical relationships among entities (e.g., lung cancer and cancer).

The entities that conform to the *is-a* blueprint were extracted and the corresponding triplets are added to OKB (e.g., < chemo, is-a, chemotherapy >, < lose weight, is-a, weight loss >, < lung cancer, is-a, cancer >. It should be noted that the head and tail entities in an *is-a* triplet are not necessarily extracted from the same post. They are created by examining entities from all posts.

Table 3.1. Example of posts from r/cancer with corresponding entities underlined in the text, relation triplets and triplet annotation. For brevity, not all possible triplets associated with the post are listed. For example, the first post has other potential triplets including < IV antibiotics, is-a-treatment-of, abscess > and < chemo, is-a-treatment-of, abscess >.

Post 1
Has anyone here contracted a nocardia infection during treatment? Hello! I'm just wondering if anyone here has contracted a nocardia infection and if so, how did this impact on their treatment? I'm currently in hospital with a nocardia infection from my Hickman line. The infection has caused a small abscess in my brain which is obviously not good. They've taken the line out and I'm on IV antibiotics. I'm getting a brain MRI on Monday to have a closer look at the abscess. I'm really worried about how this is going to impact my treatment. I still have 4 rounds of chemo left to do but it's likely I will be on antibiotics for months as this type of bacteria is very dangerous and slow to respond to antibiotics. I'll be seeing my oncologist on Monday hopefully but just thought I'd post here and see if anyone else has been in a similar situation.

Continued on the next page

Table 3.1. cont.

Blueprint	**Extracted Triplet**	**Triplet Class**
<Med./Treat./Proc., is-a-treatment-of, Diagnosis>	<IV Antibiotics, is-a-treatment-of, Nocardia infection>	Relation exists
<Diagnosis, co-existing-diagnosis, Diagnosis>	<Nocardia infection, co-existing-diagnosis, Abscess>	Relation exists
<Med./Treat./Proc, is-a-treatment-of, Diagnosis>	<Chemo, is-a-treatment-of, Nocardia infection>	No Relation
Post 2		
My friend died a few weeks ago from cancer. My friend, 19, passed away from cancer a few weeks ago and I'm still not okay. I've known her since I was 3 and we basically grew up together. Her family was my family and vice versa. She was diagnosed with two rare forms of cancer when we were 17 and died a few weeks ago after it metastasized to her brain. It's all been such a blur. I keep remembering sitting at her bedside as she was unconscious and throwing up, her family crying and rushing to clean the vomit up. I remember her funeral and touching her cold body. I've never experienced grief before, and I don't know what to do anymore because it just hurts. I just needed somewhere to get this out, and I've just been holding all of this in to stay strong for my family and hers.		
Blueprint	**Extracted Triplet**	**Triplet Class**
<Symptom/Sign, is-a-symptom-of, Diagnosis>	<Throwing up, is-a-symptom-of, Cancer>	Relation exists
<Symptom/Sign, co-existing-symptoms, Symptom/Sign>	<Unconscious, co-existing-symptoms, Throwing up>	Relation exists

Continued on the next page

Table 3.1. cont.

Post 3		
How much Colon should be removed? 31M, Large Adenocarcinoma, Splenic Flexure. Hey fellow cancer folk. I am new to the community and have a whole new level of respective for those who have been through this before, are going through it now and or who have supported others in this fight. I was diagnosed with colon cancer 4/1/2020 (not a joke, also my birthday). Pathology confirms adenocarcinoma, tumor size is 6.4cm x 6.5cm x 7.5cm and it is located just below the splenic flexure in my descending colon. Pathology also came back showing an Absent presence of PMS2, indicating a possibility of Lynch Syndrome. My surgeon (who is incredible) is suggesting a subtotal colectomy - the removal of 75% of my bowel. My question: Has anyone had a similar diagnosis/procedure and if so, what is life like afterward?		
Blueprint	**Extracted Triplet**	**Triplet Class**
<Diagnosis, co-existing-diagnosis, Diagnosis>	<Adenocarcinoma, co-existing-diagnosis, Lynch Syndrome>	Suspected Relation
<Med./Treat./Proc., is-a-treatment-of, Diagnosis>	<Subtotal Colectomy, is-a-treatment-of, Adenocarcinoma>	Relation Exists

3.2 Knowledge Base Completion Model

Triplets in the OKB are split into training, development, and testing sets. The training set is used to develop a machine learning model that is able to predict the likelihood of a new triplet belonging to the KB (i.e., being positive). The KB completion model is based on the NTN [8] architecture. In its current implementation, the vector representations are randomly generated for each word. The option to use a pre-trained language model (e.g., BERT [23]) to generate the embeddings of the words is also possible. Entities that consist of multiple words are represented by the average of the embeddings of the individual words. In this study, the embeddings have a dimension of 100 and the implementation of NTN described in [24] was used.

The OKB only contains positive triplets that have been extracted from the posts. It was therefore necessary to synthetically create negative triplets that can be used to train the KB completion model. The process of creating negative triplets is called corruption and

occurs during training [8]. Each triplet is corrupted by either replacing the head or tail with a randomly chosen entity. For example, < folfox, is-a-treatment-of, bowel cancer > may be corrupted in one of two ways:

1. Replacing the head: < melatonin, is-a-treatment-of, bowel cancer >
2. Replacing the tail: < folfox , is-a-treatment-of, urinary tract infection >

The number of corrupted triplets generated for each positive triplet during training is a hyperparameter. Moreover, the negative triplets cannot conflict with the positive triplets that already exist in the OKB since conflicting triplets can impede the learning process. Additionally, there is a constraint on the category from which the corrupted entity is sampled as it must match the category of the entity it is replacing. In the example above, melatonin was randomly chosen to replace folfox to create a negative triplet because it belonged to the same category as folfox: medication/treatment. Ideally, after training, the resulting model would score positive triplets higher than the negative ones [8].

The development set is used to determine a threshold for the score generated by the model for each triplet [8]. Triplets with scores above the threshold are classified as positive while those with scores below the threshold are classified as negative.

The testing set is used to evaluate the accuracy of the model. The same corruption process used during training is also used for the development and the testing sets with one main difference: Only one negative triplet is generated for each positive triplet in the test and development datasets [8]. This creates balanced test and development datasets with an even split of positive and negative triples. The constraint on the category of the corrupted entity matching that of the entity it replaced is still enforced in both the development and test datasets. This allows for a more rigorous testing of the inference ability of the model. It also allows for the evaluation of the ability of the model to answer questions of interest to scientists and not malformed questions. For example, a scientist may submit < ABVD, is-a-treatment-of, hodgkin's lymphoma > as a question to the model but is unlikely to ask if < fatigue, is-a-treatment-of, hodgkin's lymphoma >.

As posts may address similar topics, identical triplets may be extracted from multiple posts. For example, chemotherapy was commonly used as a treatment for cancer by users

from the sampled posts. Since the OKB includes a unique instance of these triplets, a weight is associated with each triplet based on the number of times the relation was identified in the training dataset. Table 3.2 shows the top five triplets found in the training dataset by

Table 3.2. Frequency of occurrence of triplets within the training dataset.

Triplet	Frequency of Occurrence (Weight)
< Chemo, is-a-treatment-of, Cancer >	112
< Surgery, is-a-treatment-of, Cancer >	45
< Radiation, is-a-treatment-of, Cancer >	37
< Chemo, is-a-treatment-of, Cancer >	28
< Chemotherapy, is-a-treatment-of, Cancer >	26

frequency of occurrence. As expected, the relation *is-a-treatment-of* was the only relation found among the top most frequent triplets, and the entity *cancer* always appears in the tail position of the top five triplets. This is an indication of the topic of focus in the sampled posts. The number of triplets that had a weight higher than 1 in the training set was 234. This is much lower than the number of triplets with a weight of 1 (i.e., 3,458). Triplets with higher weights are sampled more frequently during the training of the NTN model in a manner proportional to their associated weight.

4. RESULTS

The processing of the 1,000 posts randomly extracted from r/cancer resulted in a total of 1,171 medical entities. The annotation confirmed a total of 3,816 positive relations. Using the 1,171 medical entities, 800 *is-a* relations that satisfied at least one of the previously mentioned conditions were constructed. Table 4.1 shows the number of positive triplets under each relation type. Two datasets were created from the original OKB: The first dataset (baseline dataset) contains the triplets extracted from r/cancer using the process described earlier while the second dataset (enhanced dataset) also contains the same triplets in addition to the newly created is-a triplets. This enhanced dataset can allow the evaluation of the effect of adding the hierarchical relation *is-a* on the performance of the KB completion model. The triplets in both datasets were split into 80% for training, 5% for development and 15% for testing regardless of the relation type. For the baseline dataset, 3,056 positive triplets were used for training, 191 for development, and 569 for testing. For the enhanced dataset, 3,692 positive triplets were used for training, 231 for development, and 693 for testing. The difference in the dataset sizes corresponds to the number of added *is-a* triplets. After the development and splitting of the two datasets, two NTN models were developed: the first did not use weighted triplets during the training process while the second did. The former is called the baseline model while the latter is called the enhanced model.

Table 4.1. Number of positive triplets for each relation type. *Used only for the enhanced KB completion model.

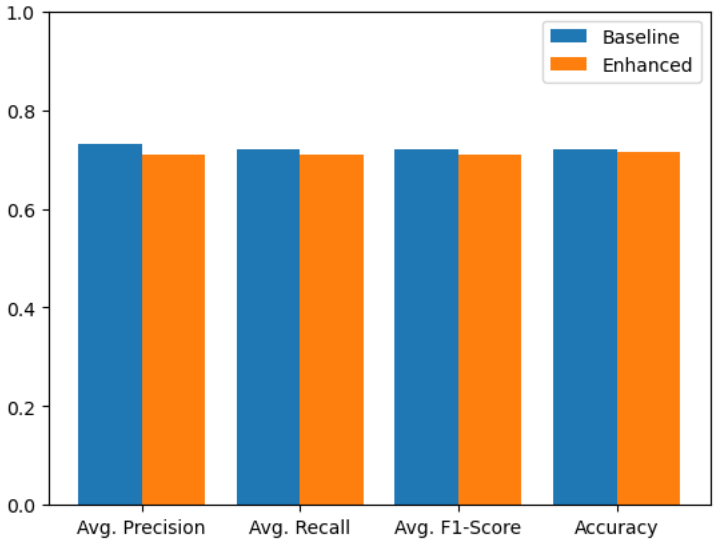
Relation	All	Training Triplets	Development Triplets	Testing Triplets	Synthetic Testing Triplets
is-a-treatment-of	1,080	878	53	149	356
is-a-symptom-of	494	409	20	65	321
is-a-side-effect-of	377	300	21	56	183
co-existing-symptoms	1,071	837	60	174	644
co-existing-diagnosis	794	632	37	125	473
is-a*	800	636	40	124	-

The baseline KB completion model is trained using the positive training triplets and negative triplets produced from the corruption process described above. The number of negative triplets generated for each positive triplet is a hyperparameter. Several values of this hyperparameter were evaluated. The results show that a ratio of three negative triplets to one positive triplet provides the best performance. Once training is completed, the baseline model is then tested using the positive testing triplets and corresponding negative triplets. One negative triplet is used for each positive triplet during testing. The performance of the model is evaluated using two metrics: accuracy and recall @k.

Table 4.2. Accuracy test results of Baseline and Enhanced models on the Baseline dataset.

	Baseline Model			**Enhanced Model**		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Positive	0.77	0.63	0.69	0.72	0.70	0.71
Negative	0.69	0.81	0.74	0.71	0.73	0.72
Accuracy		72.1%			71.4%	

Figure 4.1. Performance of Baseline and Enhanced models on the Baseline dataset.



Accuracy corresponds to the percentage of test triplets that are classified correctly by the model as either positive or negative. This metric evaluates the ability of the KB completion model to answer questions such as is Folfox a treatment of bowel cancer (<folfox, is-a-treatment-of, bowel cancer>)? Recall @k [25] evaluates the ability of the model to appropriately complete relation triplets. That is answer the question: what diseases is Folfox used to treat (<folfox, is-a-treatment-of, ?>). Based on the scoring function, the model returns a ranked list of potential tail entities that can complete the triplet. For each triplet in the test set, the percentage of times the correct tail entity is among the top k entities returned by the model is defined as the recall @k.

The enhanced KB completion model takes advantage of the availability of the *is-a* relation which is necessary for medical entities as it can help overcome variations in the expression of these entities over social media. Table 4.2 shows the performance metrics of both the Baseline and Enhanced models on the Baseline dataset. Figure 4.1 shows the average precision, recall, F1-score, and accuracy for both models. The results indicate that on the baseline dataset, the performance of both models is similar.

The same accuracy test is applied to both models on the enhanced dataset, and the results are shown in Table 4.3. The enhanced model outperforms the baseline model by 5.2% on the accuracy metric. Figure 4.2 also shows that the enhanced model outperformed the baseline model.

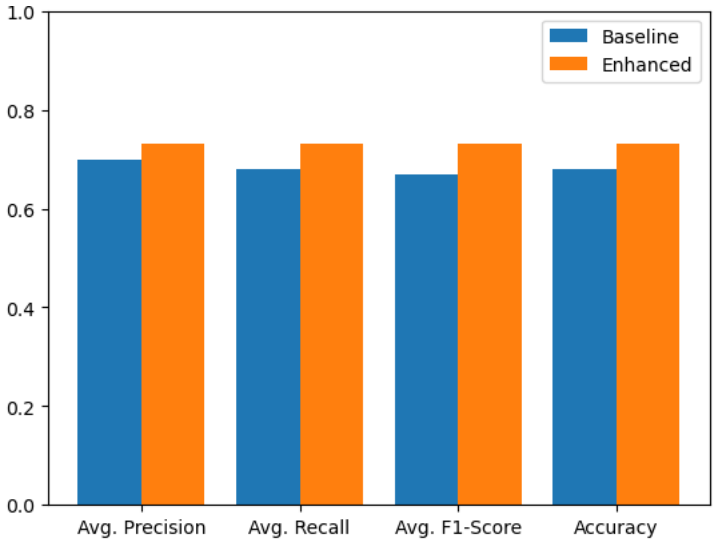
The results indicate that for both models, the precision of the positive class is always higher than that of the negative class while the recall of the negative class is always higher than that of the positive class. This result indicates that both models are being more selective when classifying a triplet as positive than when classifying a triplet as negative. Since the focus of KBs is including only true facts, this selectivity is beneficial in KB completion tasks because it reduces the number of false triplets being added to the KB.

The performance of both models was also evaluated on each individual relation. Figure 4.3 shows that the baseline model performed better on uni-directional relations while the enhanced model performed better on bi-directional ones. As expected, the enhanced model outperformed the baseline model on *is-a* triplets since it was trained to use this type of relation.

Table 4.3. Accuracy test results of Baseline and Enhanced models models on the Enhanced dataset.

	Baseline Model			Enhanced Model		
	Precision	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Positive	0.77	0.52	0.62	0.75	0.70	0.72
Negative	0.64	0.85	0.73	0.72	0.76	0.74
Accuracy		68.0%			73.2%	

Figure 4.2. Performance of Baseline and Enhanced models on the Enhanced dataset.



The second test, Recall @k, is evaluated for values of k = 10, 20, 30 and compared for the two models on the two datasets. Table 4.4 shows that the enhanced model outperforms the baseline model on both datasets. Practically, the enhanced model is able to appropriately classify and complete a triplet such as <Transarterial Chemoembolization, is-a-treatment-of, Liver Cancer> whereas this triplet is misclassified by the baseline model. Table 4.5 shows the list of top 10 entities (recall@10) returned by both models for the triplet <Transarterial Chemoembolization, is-a-treatment-of, ?>. The enhanced model returned Liver Cancer, the correct tail entity, among the top 10 possible entities that answer the question whereas for the baseline model this entity was not among the top 10.

Figure 4.3. Accuracies of Baseline and Enhanced models by individual relation.

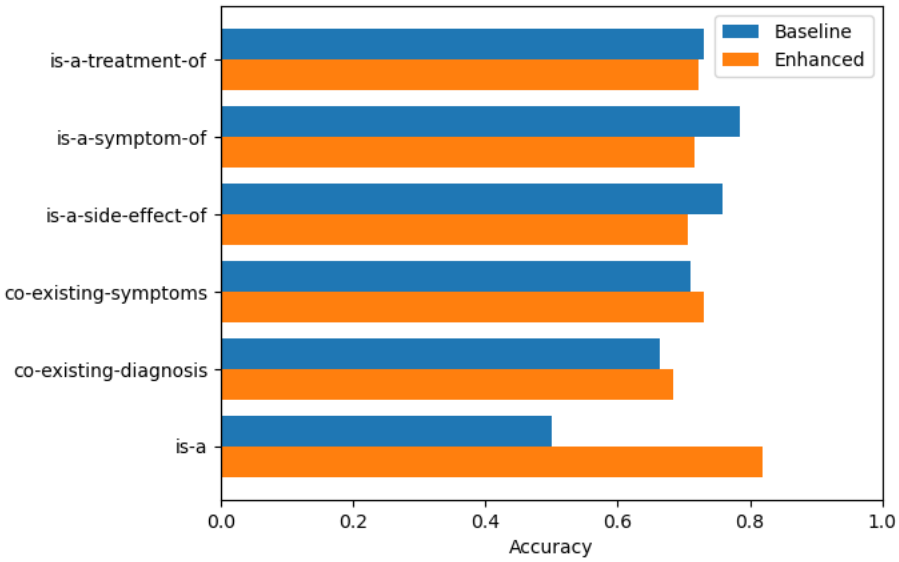


Table 4.4. Recall@k performance of Baseline and Enhanced models on both datasets.

	Recall @ K			
	Baseline Dataset		Enhanced Dataset	
K	**Baseline Model**	**Enhanced Model**	**Baseline Model**	**Enhanced Model**
10	26.7	27.2	23.5	35.4
20	35.0	39.4	31.2	46.0
30	40.6	45.7	36.5	51.5

In order to highlight the improved performance afforded by the addition of the *is-a* relation, a set of synthetic test triplets is created. The purpose of this synthetic test set is to showcase the ability of the model to perform transitive inferences which require the *is-a* relation. For example, if the triplets <diarrhea, is-a-side-effect-of, folfiri> and <folfiri, is-a, chemotherapy> are in the OKB, then a synthetic test triplet is created in the form of <diarrhea, is-a-side-effect-of, chemotherapy>, and the KB completion model is asked to classify this latter triplet as positive or negative. A total of 1,977 synthetic triplets were developed. The distribution of these triplets across the different relations is shown in Table 4.1. When tested on the synthetic set of triplets, the baseline model achieves an accuracy of 71.4%. This accuracy is lower than the accuracy achieved by the model on the test triplets

Table 4.5. Recall@10 of the Baseline and Enhanced KB models for the completion of the triplet <Transarterial Chemoembolization, is-a-treatment-of, ?>. One of the correct tail entities is Liver Cancer.

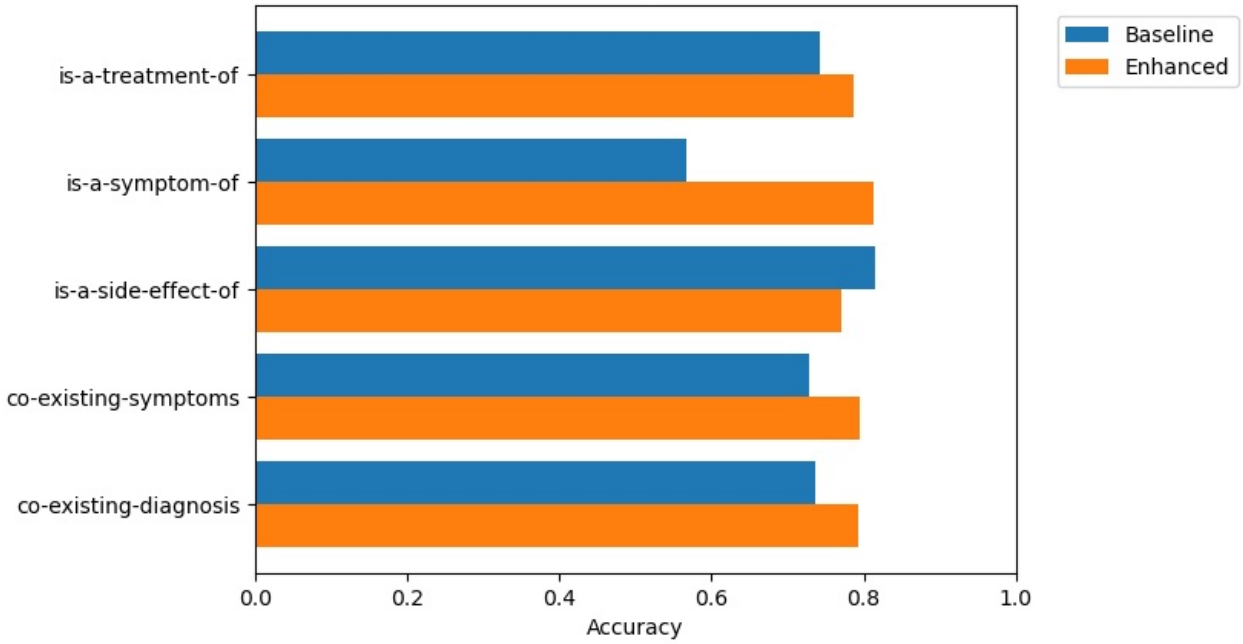
Rank	Baseline	Enhanced
1	Rash	Thyroid Cancer
2	Acne	Breast Cancer
3	Colon Cancer	Cyst
4	Depression	Precancerous Cells
5	Abscess	Cancer
6	Ascites	**Liver Cancer**
7	Thyroid Cancer	Ovarian Cyst
8	Acute Lymphoblastic Leukemia	Cervical Cancer
9	Bone Infection	Ovarian Cancer
10	Bone Cancer	Colon Cancer

directly extracted from r/cancer which was 72.1%. This result is expected since the baseline model is not trained to take advantage of the *is-a* relation. However, Table 4.6 shows that the accuracy of the enhanced model on the synthetic test triplets is 79.4%. This percentage is higher than that of the baseline model on the same dataset and even higher than the accuracy of the enhanced model on the testing triplets directly extracted from r/cancer which was 71.4%. This indicates that there was a benefit to adding the hierarchical relation triplets which improved the model's inference abilities across transitive relations. Figure 4.4 shows the performance of both models for each relation type on the transitive inference test. The enhanced model outperforms the baseline model on all the relations except *is-a-side-effect-of*. One possible explanation is that there were not enough training triplets from this particular relation as indicated in Table 4.1, or that the threshold score found was not a good separator between negative and positive triplets again as result of an insufficient number of triplets.

Table 4.6. Performance of Baseline and Enhanced models on synthetic test set.

	Baseline	**Enhanced**
Accuracy (synthetic test)	71.4%	79.4%

Figure 4.4. Accuracies of Baseline and Enhanced models on Synthetic test set by individual relation.



5. CONCLUSION

The ability to analyze the experience of patients and caregivers with chronic diseases such as cancer is important to both researchers and health providers. This thesis describes a methodology that can make this information readily accessible to all stakeholders. Specifically, the proposed framework consists of an oncology knowledge base that includes facts extracted from the online health forum r/cancer and a knowledge completion model that can infer new facts. The study shows that it is necessary to enhance the knowledge base by adding is-a relations that can help define how entities that belong to the same category are associated with one another (e.g., <folfiri, is-a, chemotherapy>).

The results also show that the framework is able to classify the five types of relations under consideration in the present study, namely *is-a-treatment-of, is-a-symptom-of, is-a-side-effect-of, co-existing-symptoms, co-existing-diagnosis*, as either positive or negative. Moreover, the framework is able to complete any of the above relations with a recall @10 of 35.4%. These two capabilities can allow scientists to answer questions such as "what are the reported side effects of chemotherapy?" according to the posts submitted by patients and caregivers to r/cancer. Scientists can also drill down and compare different chemotherapy treatments by asking questions such as "what are the reported side effects of Folfiri?" versus "what are the reported side effects of mXeliri?". In addition, scientists can use the framework to answer questions such as "what is the most discussed treatment of colon cancer?"

Although useful for research, extending the proposed framework or adapting to other chronic diseases has several challenges. One of the main challenges is the effort required to annotate the positive triplets from the posts. Entities can be identified using an automated named entity recognition algorithm and the present study defines blueprints that only generates plausible triplets. However, these triplets need to be manually labeled according to each post. Moreover, a large number of triplets requiring annotation may be extracted depending on the length of the post. Another challenge is the ability to construct *is-a* triplets for all the entities in the same category. Overcoming these two challenges is the subject of current investigation. Future work also considers implementing the proposed enhancements to other

KB inference models and developing techniques that allows relations to adapt dynamically to new contexts.

REFERENCES

[1] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: A collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp. 1247–1250.

[2] L. M. Bernstein, E. R. Siegel, and C. M. Goldstein, "The hepatitis knowledge base: A prototype information transfer system," *Annals of Internal Medicine*, vol. 93, no. 1_Part_2, pp. 169–181, 1980.

[3] H. Wang, F. Zhang, J. Wang, M. Zhao, W. Li, X. Xie, and M. Guo, "Ripplenet: Propagating user preferences on the knowledge graph for recommender systems," in *Proceedings of the 27th ACM international conference on information and knowledge management*, 2018, pp. 417–426.

[4] A. Bordes, S. Chopra, and J. Weston, "Question answering with subgraph embeddings," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 615–620. DOI: 10.3115/v1/D14-1067.

[5] Q. Wang, B. Wang, and L. Guo, "Knowledge base completion using embeddings and rules," in *Twenty-fourth international joint conference on artificial intelligence*, 2015.

[6] J. Baumgartner, S. Zannettou, B. Keegan, M. Squire, and J. Blackburn, "The pushshift reddit dataset," in *Proceedings of the international AAAI conference on web and social media*, vol. 14, 2020, pp. 830–839.

[7] N. Andalibi, O. L. Haimson, M. D. Choudhury, and A. Forte, "Social support, reciprocity, and anonymity in responses to sexual abuse disclosures on social media," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 25, no. 5, pp. 1–35, 2018.

[8] R. Socher, D. Chen, C. D. Manning, and A. Ng, "Reasoning with neural tensor networks for knowledge base completion," *Advances in neural information processing systems*, vol. 26, 2013.

[9] B. Yang, S. W.-t. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," in *Proceedings of the International Conference on Learning Representations (ICLR) 2015*, May 2015.

[10] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," *Advances in neural information processing systems*, vol. 26, 2013.

[11] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph embedding by translating on hyperplanes," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 28, 2014.

[12] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *International conference on machine learning*, PMLR, 2016, pp. 2071–2080.

[13] Y. Shen, P.-S. Huang, M.-W. Chang, and J. Gao, "Modeling large-scale structured relationships with shared memory for knowledge base completion," in *Proceedings of the 2nd Workshop on Representation Learning for NLP*, Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 57–68. DOI: 10.18653/v1/W17-2608. [Online]. Available: https://aclanthology.org/W17-2608.

[14] Y. Lin, Z. Liu, M. Sun, Y. Liu, and X. Zhu, "Learning entity and relation embeddings for knowledge graph completion," in *AAAI*, 2015.

[15] O. Bodenreider, "The unified medical language system (umls): Integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004.

[16] A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini, and V. A. McKusick, "Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders," *Nucleic acids research*, vol. 33, no. suppl_1, pp. D514–D517, 2005.

[17] M. Palmonari and P. Minervini, "Knowledge graph embeddings and explainable ai," *Knowledge Graphs for Explainable Artificial Intelligence: Foundations, Applications and Challenges*, vol. 47, p. 49, 2020.

[18] D. Q. Nguyen, "A survey of embedding models of entities and relationships for knowledge graph completion," in *Proceedings of the Graph-based Methods for Natural Language Processing (TextGraphs)*, Barcelona, Spain (Online): Association for Computational Linguistics, Dec. 2020, pp. 1–14. DOI: 10.18653/v1/2020.textgraphs-1.1.

[19] P. Bhatia, B. Celikkaya, M. Khalilia, and S. Senthivel, "Comprehend medical: A named entity recognition and relationship extraction web service," in *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*, IEEE, 2019, pp. 1844–1851.

[20] *Detect entities (version 2)*. [Online]. Available: https://docs.aws.amazon.com/comprehend-medical/latest/dev/textanalysis-entitiesv2.html (visited on 02/05/2021).

[21] U. Naseem, M. Khushi, V. Reddy, S. Rajendran, I. Razzak, and J. Kim, "Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition," in *2021 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2021, pp. 1–7.

[22] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and robust models for biomedical natural language processing," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 319–327. DOI: 10.18653/v1/W19-5034.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.

[24] Siddharth Agrawal, *Neural-tensor-network*, Sep. 21, 2014. [Online]. Available: https://github.com/siddharth-agrawal/Neural-Tensor-Network.

[25] D. Chen, R. Socher, C. D. Manning, and A. Ng, "Learning new facts from knowledge bases with neural tensor networks and semantic word vectors," *CoRR*, vol. abs/13-01.3618, 2013.

A. ANNOTATION GUIDELINES

This appendix describes the annotation guidelines given to the annotators to explain the task of labeling relation triplets given the content of a post from r/cancer.

The goal of this annotation task is to label posts from the online forum reddit.com/r/-cancer according to the existence of a relation between entities. Posts from the cancer subreddit are first collected and medical entities are extracted from the posts. Triplets are formed using the extracted entities for a predefined set of relations. A triplet consists of a head entity, a relation, and a tail entity (example 1).

Example 1: Triplet examples:

- < chemotherapy, is-a-treatment-of, cancer >
- < hair loss, is-a-side-effect-of, chemo >
- < Fever, coexisting-symptom, weight loss >

The first term is called the head entity. The second term is called the relation. The third term is called the tail entity. Three out of the five relations provided in this task are uni-directional. Two relations are bi-directional.

A.1 Relations

This annotation task is focused on a set of five relations as follows:

- **is-a-treatment-of**: Is this medication/surgery/treatment course (head) used as a treatment for this disease (tail)?
- **is-a-symptom-of**: Is this symptom (head) a symptom of this disease/diagnosis (tail)?
- **is-a-side-effect-of**: Is this symptom (head) a side effect of this treatment (tail)? It should be noted that side effects occur after taking the treatment.
- **co-existing symptom** (Bi-directional): Does the patient suffer from these two symptoms simultaneously?

- **co-existing diagnosis** (Bi-directional): Does the patient suffer from these two diseases simultaneously?

The task is to label whether the two entities contained within the triplet exhibit the given relation solely based on the content of the post.

There are two types of bi-directional relations: co-existing symptoms and co-existing diagnosis. The direction from which the entities forming a triplet containing a bi-directional relation is read does not matter (example 2).

Example 2: < lung cancer, co-existing diagnosis, diabetes > should be treated the same way as < diabetes, co-existing diagnosis, lung cancer >.

A.2 Blueprint of Relations

Each relation has a predefined blueprint from which it is constructed. Triplets are formed using comprehend medical [19] which identifies the general medical categories [20] of extracted triplets e.g., Chemotherapy would be categorized as a treatment while cancer would be a diagnosis. The following are the blueprints of categories for each relation:

- **Is-a-treatment-of**: < Treatment/Medication/Procedure, is-a-treatment-of, Diagnosis >
- **is-a-side-effect-of**: < Symptom/Sign, is-a-side-effect-of, Treatment/Medication/Procedure >
- **is-a-symptom-of**: < Symptom/Sign, is-a-symptom-of, Diagnosis >
- **co-existing symptoms**: < Symptom/Sign, co-existing symptoms, Symptom/Sign >
- **co-existing diagnosis**: < Diagnosis, co-existing diagnosis, Diagnosis >

A.3 Classes

A triplet can fall under one of the three following classes:

- **Relation Exists**: The head and tail entities exhibit the given relation based on the accompanying post. There can be direct evidence from the post supporting this or it can be inferred based on the information given in the post.
- **No Relation**: There is no evidence from the accompanying post that the head and tail entities exhibit the given relation.
- **Suspected Relation**: There is evidence that suggests that the poster/patient/physician/caregiver suspects that this relation exists. However, it has not been confirmed.

The following are three extra considerations that are needed in order to maintain the quality of the annotated data. Effort was taken in order to manually reduce instances of triplets that exhibit the following characteristics. However, triplets with similar characteristics may remain.

1. The entities in this task were extracted using an automated process. Some of them do not make sense as specific medical entities. Others are phrases that should not be considered as medical entities. When such a triplet is encountered, it should be labeled no relation. Note that at least one out of the two entities can satisfy this condition for the triplet to be labeled no relation. (example 8)
2. Triplets will sometimes contain a specific type of cancer as one entity and cancer as the second entity e.g., (brain cancer, co-existing diagnosis, cancer). These triplets should also be labeled no relation. This occurs mostly when the relation is co-existing diagnosis (example 9).
3. Triplets will sometimes contain variations of the same entity, with one as a head and the other as a tail. These triplets should be labeled no relation (example 10).

A.4 Specific Examples

This section includes several examples of labelling extracted triplets. In addition, it provides an explanation of the reasoning behind choosing a particular classification

A.4.1 Relation Exists Classification

Example 3: "Sleep problems Recently Ive had problems with sleep. Some real problems. Ive had times when I couldnt sleep, it would take me a long time to fall asleep and when I do I sleep way too much. Ive had times where Ive been awake for over 24 hours and cant sleep, on the other hand when I do sleep, I sleep for 14-20 hours. Even if I havent been awake for a really long time Ill sleep far too much. Also during those times, I will wake myself up by hitting myself in the head. On the side of my craniotomy I had in 2016. Sometimes rubbing hard or scratching, but doing something while sleeping. Im stage 4 melanoma. I had a neck dissection in 2015 to remove tumor in my neck, in 2016 craniotomy to remove 4cm tumor in left frontal lobe over Brocas area. In 2015 I was on ipilumumab and keytruda after craniotomy. Radiation treatments for both of course. Im wondering what could be causing this? Im going to ask my oncologist of course, Im just wondering if anybody has had this happen. Its very strange to wake up sitting up and hitting yourself."
Sample extracted triplet: < 'keytruda', 'is-a-treatment-of', 'melanoma' >
From the post, we can tell that the patient has taken Keytruda to treat melanoma. Therefore, the triplet is true, and the label would be Relation Exists

Example 4: "I'm a mess On November 27th at 7 in the morning my sister passed.I'm still in shock and disbelief from it, but slowly getting there.It was so fast. We had seen her on wednesday, she was happy, talking and in a good mood. Friday night she had a massive headache, was throwing up and not doing well. Saturday morning, 911 was called and she became non-responsive. Saturday night they moved her to palliative care. Monday morning, she passed.We're all still dealing with it and I'm just so glad I cancelled my meeting on that wednesday to go see her. If I hadn't I would have never been able to talk to her or tell her I love her one last time.I just needed to vent and thank everyone here for the help they gave in the past."

Sample extracted triplets: <'throwing up', 'co-existing-symptoms', 'headache'>
The patient experienced "throwing up'" and "headache" around the same time frame. Therefore, the triplet is true, and the label would be Relation Exists.

Example 5: "Will my chemo affect any future children in any way? I went through testicular cancer early this year and after removal and chemo to treat I am cancer free. I had banked sperm just incase but since finishing my treatment i have been tested and am still fertile. (very low in mobility and count)My partner is understandably of thrilled with the idea of IVF and my question is if we where to consive naturally now as apposed to doing IVF with my pre chemo banked sperm would it affect how our child may turn out? I.e is there a greater chance of something being wrong with the child one way or another?"
Sample extracted triplet: < 'chemo', 'is-a-treatment-of', 'testicular cancer' >
From the post, we can tell that chemo was used by patient to treat testicular cancer. Therefore, the triplet is true, and the label would be Relation Exists.

A.4.2 Relation Does Not Exist Classification

Example 6: "Recovery from lobectomy Hi all, I had stage 1B lung cancer, treatment was lobectomy of my left upper lobe. I have had an extremely slow recovery from surgery. I'm still having a significant amount of pain and I'm not able to do much physically after 6 months. I talked to my surgeon and his assistant many times. Their answer was to give me pain meds and eventually ignore me. I went to my PCP and told her my concerns and she gave me more pain meds but she also referred me to physical therapy. It has been a mixed blessing. I do think it's helping. However, it is extremely painful. They give me some exercises to do but mainly I get deep tissue massage-they tell me I have scar tissue and that my muscles have basically atrophied so it will take lots of work to get me back to where I was. I also have a significant amount of nerve damage. I know from prior surgery that it can take years for nerve pain to abate. I am venting mostly, but does anyone else have experience with this? I'm wondering if I will need to go on pain management."
Sample extracted triplet: < 'pain', 'is-a-side-effect-of', 'physical therapy' >

There's no evidence that suggests pain is a side effect of physical therapy in the post. Therefore, the relation does not exist, and the label assigned would be No Relation.

Example 7: "Biopsy after finishing chemo Hi r/cancerI just had my post-chemo scans and was told that mostly everything looks better but there is an ambiguous, mildly hypermetabolic area that showed up on the PET scan that my oncologist is not sure if it's remaining cancer or an infection or inflammation. I'm a little shocked because, while I knew this was a possibility, after surgery, radiation, and chemotherapy, I thought I would be fine. My oncologist is known to be extremely cautious about everything and she was telling me that there's a possibility that I'd have to have another biopsy of the suspicious area. Has this happened to anyone else before? If it was malignant, would I have to do even more chemo? This seems crazy to me. I mean, I know I didn't have the best luck with getting cancer in general but after having gone through nearly 9 months of treatment, I thought I'd paid my dues. Maybe I'm overreacting. I guess I'm just overwhelmed."
Sample extracted triplet: < 'chemo', 'is-a-treatment-of', 'inflammation' >
There's no evidence that chemo was used as a treatment for inflammation in the post. Therefore, the relation does not exist, and the label would be No Relation.

Example 8: "Votrient side effects My brother has been battling DSRCT for 18 months and we just experienced a huge blow in the form of side effects from the votrient. We thought he had developed pneumonia and he was hospitalized last week. His lung capacity is severely diminished and he struggles to breath, even on oxygen. All tests came back negative for fungal, bacterial, viral, and even cancer progression. The doctors feel it is an uncommon side effect from the chemotherapy. Does anyone have any first hand experience with this? Will his lungs heal? Is it temporary? They're trying steroids now but it has not helped. His only chance to prolong his life is through surgery and he can not do it in his current condition. I'm getting desperate for anything information wise. I don't want this to end like this."
Sample extracted triplet: < 'side effects', 'is-a-symptom-of', 'battling dsrct' >
"Battling dsrct" is not a medical entity. It does not refer to a symptom or a medical condition but rather to the act of battling one. "Side effects" does not refer to a specific medical

entity. This example satisfies condition 1 mentioned in Section A.3, therefore this triplet should be labeled No Relation.

Example 9: "21f, likely Hodgkins Lymphoma recurrence i dont know where else to post thisMy friends and family understand, but they dont UNDERSTANDIm 21, and a little over a year ago was my last treatment for stage 4 hodgkinsim showing signs of recurrance,my doctors and family are worried, i took a blood test today and theyre going to call me to schedule my PETi had to use my inhaler for the first time in a year last night after just going to the grocery store, and today i nearly collapsed making dinnerim 21,single,live alone with my cats, and tried to commit suicide 6 months ago, and now my cancer is probably backi justim on the verge of a meltdown"
Sample extracted triplet: < 'cancer', 'co-existing-diagnosis', 'hodgkins lymphoma'>
Hodgkins lymphoma is a type of cancer. This satisfies condition 2 mentioned in Section A.3, therefore this triplet should be labeled No Relation.

Example 10: "Throat Cancer and sense of smell heightened need advice My father is entering week 5 of treatments and taste/saliva almost gone as expected. He is complaining that his sense of smell is so heightened that it is really bothering him and even making him sick. Anyone know any potential remedies to help with this? "
Sample extracted triplet: < 'sense of smell heightened', 'co-existing-symptoms', 'sense of smell is so heightened' >
The head and the tail entities are variations of the same entity. This triplet satisfies condition 3 mentioned in Section A.3, therefore this triplet should be labeled No Relation.

A.4.3 Suspected Relation Classification

Example 11: "What to expect/questions to ask? So, I've had a lump under my tongue for around a year and a half. It's quite large and rounded, about the size of a child's marble. Up until I gave up smoking and drinking last year I regularly found blood in my saliva with no obvious source - apart from generally feeling very tired and unwell those were my only

symptoms.I finally went to the doctor last week, and she was able to feel the lump from underneath my jawline, and referred me for an ultrasound. I've translated the documents that I'm supposed to give the ultrasound clinic into english (I'm a brit but I live abroad, the documents are in portuguese) and it seems that the doctor suspects a submandibular salivary gland neoplasm and has asked them to investigate.I'd be grateful if someone could help me with a few questions that have been on my mind since then:- What is she looking for with the ultrasound? are there benign explanations for the swelling?- What further tests might be needed? a biopsy?- What exactly is a biopsy? how much tissue gets removed?Thanks in advance."

Sample extracted triplet: < 'tired', 'is-a-symptom-of', 'salivary gland neoplasm' >
It is not confirmed that the patient has salivary gland neoplasm. The doctor only suspects it. Therefore, this triplet should be labeled suspected relation.

Example 12: "Lump and pain on right side on neck for about a year now? So about a year ago, I notice a small 1cm lump that was a white ish colour on the right side of my throat just below the right tonsil. Since then, there has been a pain constantly going on and off about there, so I suspect that it is coming from the lump. I have been to a GP many times for colds and such, and they have never said anything about it when they look at my throat. One day I feel the pain, and one day I don't it had been like this for ages. Also, I think that my lymph nodes are slightly swollen or something, from around the same time, although the doctor never said anything about swollen glands either. The doctors say they are fine. I am recently freaking about this, as those three things, the lump, pain, and slight swelling might be symptoms of cancer. I've never really had any other cancer symptoms, the lump is not bigger, it stays the same but its been the same for a year so I'm worried any help?"

Sample extracted triplet: <'pain', 'is-a-symptom-of', 'cancer' >
There's no evidence that the patient has cancer from the post. The patient is only worried that pain may be a symptom of a cancer. Therefore, this triplet should be labeled suspected relation.