Embedding Open-domain Common-sense Knowledge from Text

Travis Goodwin and Sanda Harabagiu
Human Language Technology Research Institute
University of Texas at Dallas
Richardson, TX 75083-0688, USA
{travis,sanda}@hlt.utdallas.edu

Abstract

Our ability to understand language often relies on common-sense knowledge – background information the speaker can assume is known by the reader. Similarly, our comprehension of the language used in complex domains relies on access to domain-specific knowledge. Capturing common-sense and domain-specific knowledge can be achieved by taking advantage of recent advances in open information extraction (IE) techniques and, more importantly, of knowledge embeddings, which are multi-dimensional representations of concepts and relations. Building a knowledge graph for representing common-sense knowledge in which concepts discerned from noun phrases are cast as vertices and lexicalized relations are cast as edges leads to learning the embeddings of common-sense knowledge accounting for semantic compositionality as well as implied knowledge. Common-sense knowledge is acquired from a vast collection of blogs and books as well as from WordNet. Similarly, medical knowledge is learned from two large sets of electronic health records. The evaluation results of these two forms of knowledge are promising: the same knowledge acquisition methodology based on learning knowledge embeddings works well both for common-sense knowledge and for medical knowledge Interestingly, the common-sense knowledge that we have acquired was evaluated as being less neutral than than the medical knowledge, as it often reflected the opinion of the knowledge utterer. In addition, the acquired medical knowledge was evaluated as more plausible than the common-sense knowledge, reflecting the complexity of acquiring common-sense knowledge due to the pragmatics and economicity of language.

Keywords: Common-sense knowledge, medical domain knowledge, knowledge embedding

1. Introduction

Our ability to understand language often relies on common-sense knowledge – background information the speaker can assume is known by the reader. When language is used to communicate, common-sense knowledge is not articulated in most of the cases. Hence, the availability of resources capturing or approximating common-sense knowledge is crucial. Multiple attempts have been made to capture general common-sense knowledge, such as the CYC project (Lenat and Guha, 1989) and ConceptNet (Speer and Havasi, 2013). Moreover, WordNet (Miller, 1995) captures lexico-semantic knowledge in English, while DBPedia (Auer et al., 2007) provides knowledge that was used for common-sense reasoning. These resources encode a large number of *concepts* without representing all their possible attributes. Moreover, another important limitation of these existing resources for common-sense knowledge stems from the limited number of relation types spanning concepts. Some of the existing relations are taxonomic, e.g. IS-A, PART-OF, while others capture causality, e.g. ENABLE, CAUSE-OF, but they hardly represent all relation types we use when accessing common-sense knowledge. To address these two limitations, we took advantage of recent advances in open information extraction (IE) techniques and, more importantly, of the advent of *knowledge embeddings*, which are multi-dimensional representations of concepts and relations.

The use of knowledge embedding enabled us to consider *lexicalized relations* between concepts, allowing relational similarity to be easily identified, even if not grouped in the same type of relations. Although this is an approximation of the relation types available from existing common-sense knowledge resources, it captures a much larger set of possible relations spanning concepts. Because concepts are largely expressed as noun phrases, in our formulation based on knowledge embeddings, we make use of semantic compositionality to represent complex concepts that comprise their attributes and their nominal interpretation (e.g. "plantain slices" are interpreted as slices of plantain, while "plantain dish" can be interpreted as a dish made with plantains). When capturing common-sense knowledge from a variety of sources, including books and weblogs, we rely on open-domain information extraction (IE) to identify concepts as noun phrases and the lexicalized relations they share. For example, consider the following sentence obtained from a personal weblog post:

$$\overbrace{\textit{plantain slices}}^{c_1}\ \underbrace{\textit{are cooked in}}_{r}\ \overbrace{\textit{hot oil}}^{c_2}$$

This sentence encodes the lexical relation $r =$ *are cooked in*, indicating the relationship between two concepts $c_1 =$ *plantain slices*, and $c_2 =$ *hot oil*. Representing the knowledge encoded in this sentence requires a level of granularity not supported by existing knowledge sources: we must restrict ourselves to *slices of plantains* (not, for example, *slices of pie*) and to *hot oil* (not *cold oil*, and certainly not *crude oil*). Additionally, the lexical relation r between these two concepts is not present in the fixed set of relation types used by any of these knowledge sources, and cannot be approximated by the fixed set of relation types without adjusting the semantics of the original sentence. Capturing common-sense knowledge is enabled by organizing the extracted concepts and their lexical relations into a knowledge graph which we embed in a multi-dimensional vector space.

Because our approach for capturing knowledge is able to identify relational knowledge which is not currently available in existing resources, we explored the possibility of

using the same knowledge acquisition methodology for capturing domain-specific knowledge that is not current available despite domain-specific texts being abundant. For this purpose, we considered the domain of medicine.

In the medical domain, multiple efforts have been made to capture knowledge, including the Unified Medical Language System (UMLS) (Bodenreider, 2004) and the Systemized Nomenclature for MEdicine (SNOMED) (Stearns et al., 2001). These medical knowledge resources have similar limitations as the common-sense knowledge resources, namely (1) too few types of relations between concepts and (2) description of concepts without capturing their semantic concepts. However, when processing medical language, a large variety of relations need to be considered. For example, given a sentence stating:

$$\overbrace{\textit{abnormal EEG}}^{c_1} \underbrace{\textit{due to}}_{r} \overbrace{\textit{rhythmic background slowing}}^{c_2}$$

the causality between the "abnormal" attribute of the electroencephalogram (EEG) and the rhythmic slowing of the background signal needs to be recognized. Using knowledge embeddings to encode medical knowledge provides access to medical information not available in current ontologies but expressed in medical texts.

An important aspect of the common-sense and domain-specific knowledge acquisition framework described in this paper is provided by our probabilistic treatment of relations between concepts. By encoding the *plausibility* of a lexicalized relation between a pair of concepts, e.g. $P(q) = \langle c_1, r, c_2 \rangle$, the knowledge encoding framework that we describe is able to capture salient information as knowledge embeddings without discarding the long-tail aspects of this knowledge. Moreover these knowledge embeddings are able to generalize the knowledge discerned from individual sentences by incorporating two notions of semantic smoothness: (1) *functional similarity* (Turney, 2012), the idea that two phrases are similar if they engage in similar lexical relations, and (2) *behavioral similarity* (Nakov and Hearst, 2008), the idea that two lexical relationships are similar if they operate on similar concepts.

In this paper, we present a novel knowledge embedding framework that generates promising results both on common-sense knowledge and domain-specific knowledge by learning an optimal embedding for each concept and lexical relation according to (1) the functional similarity between pairs of concepts, (2) the behavioral similarity between pairs of lexical relations, and (3) the role of the semantic composition for each word occurring in each concept and lexical relation. The remainder of this paper is organized as follows. Section 2 reviews related work and Section 3 describes the datasets we considered as well as the natural language processing techniques performed on these datasets. Section 4 details our knowledge embedding framework, while Section 5 discusses our experiments, and Section 6 summarizes the conclusions.

2. Related Work

One of the most popular sources of lexico-semantic knowledge and an approximation of common-sense knowledge is WordNet (Miller, 1995) which encodes 117,000 concepts organized as synonym sets or *synsets* spanned by a small number of relation types. The relations between concepts capture hypernymy, meronymy (for nouns), entailment, causality (between verbs), as well as antonomy. In addition, each WordNet concept is associated with a *gloss* defining the concept.

While the WordNet semantic graph was carefully generated by expert lexicographers, the JeuxDeMots project (Lafourcade, 2007) provides a crowd-sourced lexico-semantic graph which captures some of the relation types encoded in WordNet (e.g. synonym, antonymy) for French. These relations were populated by allowing people to play word association games.

ConceptNet (Speer and Havasi, 2013) was also obtained through crowd-sourcing. It encodes 21 relation types which attempt to capture common-sense knowledge. However, these relations are sometimes under-specified, e.g. for the CAPABLE-OF relation, the sentence provided by ConceptNet, "Mary broke a vase" was encoded as $\langle \textit{Mary}, \text{CAPABLEOF}, \textit{vase} \rangle$. When interpreting this relation, what does it mean to be *capable of* a "vase"? Can Mary eat a vase? Can she buy a vase? Can she cook a vase? What if we consider, instead, the relation $\langle \textit{Mary}, \text{CAPABLEOF}, \textit{break} \rangle$. Although perhaps closer to the original semantics of the sentence, it is still too general: What is Mary capable of breaking? Can she break a brick wall? Can she break a table? It would be preferable to encode only the fact that Mary broke a *vase*. However, because the relation CAPABLE-OF cannot account for the semantics of the verb "break", it generates an underspecified relation between Mary and the vase.

DBPedia (Auer et al., 2007) curates a variety of knowledge mined from Wikipedia. It considers article titles as concepts and relies on the relations between them discerned from meta-data in Wikipedia. However, the number of concepts and relation types is limited and dependent on the availability of metadata.

Medical knowledge acquisition was targeted in multiple projects. Perhaps the most popular of them is the Unified Medical Language System (UMLS) (Bodenreider, 2004) which designed by the National Library of Medicine in order to provide a standard set of concepts allowing for commonalities amongst various medical terminologies to be interlinked, for example, by mapping ICD-9 diagnostic codes (Cimino et al., 1993) to Medical Subject Headings (MeSH) (Lowe and Barnett, 1994).

A component of UMLS, the Systemized NOmenclature of MEDicine (SNOMED) (Stearns et al., 2001) provides a small number of relationship types between a subset of UMLS concepts, such as hypernymy, synonym as well as medical relations (e.g. treatment targets disease).

In addition, the Open Biomedical Ontologies (OBO) Foundry facilitates a collaborative experiment involving scientific (e.g. medical) ontologies (Smith et al., 2007).

Both the UMLS and the ontologies linked in the OBO Foundary do not represent medical knowledge probabilistically, nor do they capture many of the relations between medical concepts. Attempts to address these limitations was tackled in our previous work (Goodwin and Harabagiu,

2013) where a qualified medical knowledge graph (QMKG) was presented. However, the QMKG does not capture compositional semantics of medical concepts, nor the lexical relations between them. Both of these aspects are accounted for in the knowledge framework presented in this paper.

3. The Textual Data

To acquire common-sense knowledge as well as medical knowledge, we have relied on several sources of textual data.

3.1. Textual Data expressing Common-sense Knowledge

We considered two collections of narrative texts which we believed would be useful for common-sense reasoning: (1) narrative personal stories from weblog articles, and (2) Google's syntactic n-gram dataset obtained from stories sourced from Google Books. Narratives of personal stories have held interest within the artificial intelligence community for decades due to the rich semantic information they contain. For example, as conjectured by Schank in (Schank, 1983), personal stories can be used to pin-point gaps in a knowledge base. Moreover, as argued in (Schank and Abelson, 1995), the most common source of common-sense knowledge for humans is that of narrative communication, or stories. To provide additional common-sense knowledge, we also considered WordNet glosses.

3.1.1. Narrative Blog Posts

We used the data provided for the 2009 International Conference on Weblogs and Social Media (ICWSM), sponsored by the Association for the Advancement of Artificial Intelligence (AAAI) (Burton et al., 2009), by Spinn3r, Inc, which consists of 44 million blog posts authored between August 1st and October 1st in 2008. The dataset covers many major news events, such as the 2008 Olympics, both United States presidential nominating conventions, the United States financial crisis, etc. Gordon and Swanson identified nearly one million English-language narrative blog posts discussing personal stories from this dataset, using a supervised neural network based on lexical features (Gordon and Swanson, 2009). A total of 937,994 blog posts were classified as containing narrative stories.

3.1.2. Google Books

Inspired by (Akbik and Michael, 2014), we considered a secondary source of narrative information: the Syntactic N-Grams dataset (Goldberg and Orwant, 2013) which was extracted from 3.5 million digitized English Books available through Google Books (Michel et al., 2011). The dataset contains over 10 billion *syntactic n-grams*, which are rooted syntactic dependency tree fragments (noun phrases and verb phrases) and is the largest publicly available corpus of its kind. Each tree fragment is annotated with the dependency information, its head word, and the frequency with which it occurred. An example tree fragment is (n_0) "when/WRB the/DT small/JJ boy/NN ate/VBD cookies/NN". We converted each tree fragment in this dataset into a sentence by removing the leading relative pronouns, conjunctives, and prepositions. In this way, for the tree fragment n_0 we obtain the sentence s_0, "the small boy ate cookies."

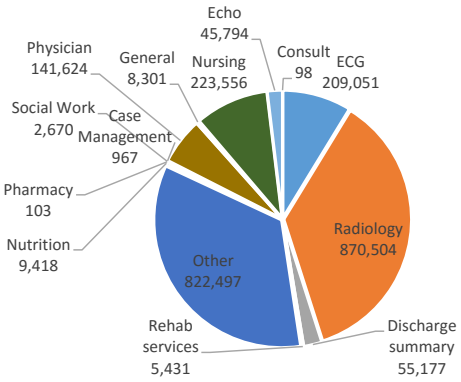

Figure 1: The distribution of electronic health record (EHR) types in the MIMIC-III clinical database.

3.1.3. WordNet Glosses

Because each concept in WordNet is (a) represented by a synset and (b) defined by a gloss, we generate a set of sentences expressing these definitions. For example, the synset $y_1 = \{$"male child", "boy"$\}$ is defined as "a youthful male person." In addition, the glosses also provide sentence examples of the concepts represented by the synset. For example, y_1 is exemplified by (s_1) "he baby was a boy," (s_2) "she made the boy brush his teeth every night", and (s_3) "most soldiers are only boys in uniform." We also created sentences based on the glosses by using each member of the synset as a genus and the gloss as the differentia. In this way, we obtain from the gloss of y_1 the sentences (s_4) "a male child is a youthful male person", and (s_5) "a boy is a youthful male person."

3.2. Medical Textual Data

With the advent of electronic health records (EHRs), a growing set of medical narratives are becoming available. These narratives reflect domain-specific medical knowledge communicated by physicians for the interpretation of other medical professionals. Thus, they reflect highly-specialized medical knowledge, both in explicit and implicit ways. We were interested to acquire medical knowledge from type of textual data: (1) medical narratives from a variety of medical record types as well as (2) EHRs which belong to a single record type. Both types of textual data needed to be available from massive medical archives.

3.2.1. MIMIC-3 Clinical Database

MIMIC-III (Medical Information Mart for Intensive Care III) is a large, freely-available database comprising de-identified clinical data for over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012 (Saeed et al., 2011). The database includes narratives of a variety of EHR types such as discharge summaries, nursing notes, and radiology reports. In this work, we considered all EHRs types, encompassing 2,426,930 medical records. The distribution of EHR types in this dataset is shown in Figure 1

3.2.2. EEG Reports

We considered the Temple University Hospital (TUH) corpus of Electroencephalogram (EEG) reports (Harati et al.,

2014). Generated by the The Neural Engineering Data Consortium (NEDC) at Temple University, the corpus contains nearly 20,000 EEG reports conducted from 2002 to 2013. The reports include the patient's clinical history and medications, a description of the EEG setting and configuration, as well as the physician's *impressions* and *clinical findings*. The descriptions document any notable epileptic activity observed by the physician when interpreting the EEG signal data, such as:

> The EEG is diffusely slow but the patient transitions in and out of a drowsy state with varying amounts of beta. There is small, shifting asymmetries noted. Stimulation of the patient produces eye opening.

The neurologist also records her impressions of the previously described observations and epileptiform activities, e.g.:

> Abnormal EEG due to:
> Generalized background slowing.
> Shifting slowing.
> Shifting beta asymmetries.

The EEG impression is further analyzed for possible *clinical correlations* or *clinical findings*, in which the physician interprets the findings and describes a diagnosis or general conclusion, e.g.:

> No epileptiform features were observed.
> This EEG is more supportive of a bihemispheric process.

4. Knowledge Acquisition Framework

To generate knowledge embeddings as representations of (1) common-sense knowledge as well as (2) medical domain knowledge, we used the following steps:
STEP 1: *Extract* concepts and the lexicalized relations between them from texts;
STEP 2: *Generate* a knowledge graph representing the concepts as vertices and the lexicalized relations as edges between them; and
STEP 3: *Learn* an optimal, semantically-smooth embedding of this knowledge graph. This knowledge acquisition framework has the advantage of capturing all the lexicalized relations in which a concept is involved. In addition, it learns representations which involve compositional semantics both at concept and relation levels. This representation is further enhanced through smoothing which resolves the data sparsity problem. Finally, the knowledge embedding which is learned informs the plausibility estimation of both existing and new relations between concepts.

4.1. Information Extraction of Concepts and Relations

Stanford's automatic open-domain information extraction system (Angeli et al., 2015) enabled us to discover *structured relation triples* (SRTs) from natural language sentences. SRTs were discovered without relying on a predefined set of concepts, set of relation types, or vocabulary. Figure 2 illustrates examples of SRTs discovered from personal webplogs, Google Books, as well as WordNet glosses. Figure 3 illustrates examples of SRTs extracted from the MIMIC-III database as well as the TUH EEG corpus.

General Knowledge
Personal Story Weblogs
⟨*white phosphorus smoke*, *may cause*, *burns*⟩
⟨*smoke*, *is*, *mist*⟩
⟨*plantain slices*, *are cooked in*, *hot oil*⟩
⟨*plantains*, *are immediately placed in*, *salt water*⟩
⟨*scattered showers*, *caused*, *flooding in parts of the state*⟩
⟨*urban areas*, *were hit by*, *the deluge*⟩
Google N-Grams
⟨*bananas*, *hung above*, *assortment of fruit*⟩
⟨*smoke*, *rose above*, *the trees*⟩
WordNet
⟨*plantains*, *are*, *starchy banana-like fruit*⟩
⟨*the fire*, *produced*, *tower of black smoke*⟩
⟨*the plains*, *are fertilized by*, *annual floods*⟩

Figure 2: Examples of structured relation triples expressing common-sense knowledge.

Medical Knowledge
MIMIC III Notes
⟨*large hematoma*, *caused*, *outlet obstruction*⟩
⟨*aneurysm*, *was coiled with*, *4-mm 360 coils*⟩
⟨*aneurysm*, *obliterated by*, *successive GDC coils*⟩
EEG Reports
⟨*drowsiness*, *is characterized by*, *increase in background beta*⟩
⟨*abnormal EEG*, *due to*, *left anterior temporal sharp wave*⟩
⟨*excess theta*, *may be due to*, *PRES syndrome*⟩
⟨*generalized spike & wave*, *consistent with*, *generalized epilepsy*⟩
⟨*mildly abnormal EEG*, *due to*, *mild background slowing*⟩
⟨*abnormal EEG*, *due to*, *rhythmic background slowing*⟩

Figure 3: Examples of structured relation triples expressing medical domain knowledge.

4.2. Generating the Knowledge Graph

In order to compactly represent the semantic information obtained from each SRT extracted from each sentence in our dataset, we generated a knowledge graph in which each concept was represented by a vertex, and each lexical relation between two concepts by a directed edge. Because we considered lexical relations between concepts, rather than individual words, the resulting knowledge graph was quite large, containing 32,522,807 vertices and 52,209,411 edges. Thus, we relied on the Apache Spark (Zaharia et al., 2010) web-scale data processing engine to generate a distributed knowledge graph. This allowed us to leverage Spark's GraphX library (Xin et al., 2013) for graph-parallel computations. However, processing such a large knowledge graph is challenging, even in a distributed architecture. Hence, we decided to reduce the number of edges without information loss. This was achieved by replacing all edges corresponding to each extraction of the same SRT with a single edge which also encodes the frequency of the ST in the dataset. We implemented this by a *map-reduce* counting operation. However, this did not account for another problem which arose during knowledge acquisition: knowledge sparsity. Because concepts are represented along with their attributes and lexicalized relations do not provide means of capturing the similarity of meaning, we do not recognize in data, no matter how large, important relational information that accounts for common-sense knowledge or domain specific knowledge. For example, in the knowledge graph,

it is difficult to recognize that the concept *hot oil*, *cooking oil*, and *olive oil* may be similar despite each corresponding to separate nodes. Likewise, each lexical relation is represented as a separate edge, such that it is difficult to identify similar lexical relations, e.g. *are cooked in*, *are cooked with*, or *were cooked in* each constitute different edges in the knowledge graph. In order to address these problems, we learned an optimal *embedding* of our knowledge graph into a semantically-smooth continuous vector space. Moreover, this embedding allowed us to infer non-explicit relational knowledge.

4.3. Learning the Knowledge Embedding

Recently, new methods for knowledge graph completion have produced a promising paradigm for embedding which is able to infer new knowledge from a knowledge graph (Bordes et al., 2013). In this paradigm, each concept is represented as a N-dimensional vector and each relation is represented by an operation in the $\mathbb{R}^N$ space such that new knowledge can be asserted by simple vector operations. The embeddings are learned by minimizing a global loss function over all the concepts and their relations in the knowledge graph. However, new knowledge often contains concepts which are not present in the knowledge graph, a problem which we address through the semantic compositionality of concept and relation embeddings. Moreover, smoothing these enhanced embeddings produces additional new knowledge. Given these observations, learning the knowledge embedding is performed in five steps:

LEARNING STEP 1: Multi-dimensional representation of concepts and relations;

LEARNING STEP 2: Accounting for semantic composition;

LEARNING STEP 3: Evaluating the plausibility of knowledge;

LEARNING STEP 4: Optimizing knowledge plausibility; and

LEARNING STEP 5: Smoothing the knowledge embeddings.

4.3.1. Multi-dimensional Representation of Concepts and Relations

As in TransE (Bordes et al., 2013), each vertex representing a concept in the knowledge graph was cast as a point in the continuous space of $\mathbb{R}^N$, where N is a parameter indicating the cardinality of our vector representation (in our case, $N = 200$). Likewise, each edge representing a lexical relation is interpreted as the *translation vector* which connects the points representing its arguments. Unlike TransE, however, we also learned a vector for each word used to express every concept as well as each word used to express the lexical relation. This allowed our embedding space to account for the role of semantic composition on (1) the types of lexical relations shared by concepts, and (2) the types of concepts spanned by each lexical relation.

Because concepts and relations are often expressed by more than one word, the embedding needs to represent the semantics of the entire expression. Single words are cast as points in $\mathbb{R}^N$. It happens that noun phrases (which account for concepts) and verb phrases (which account for relations) havbe a syntactic head which corresponds to a single word. Thus, the entire concept or lexical relation is viewed as a modification of the head words provided by the role of the modifiers in the noun phrase or verb phrase, respectively.

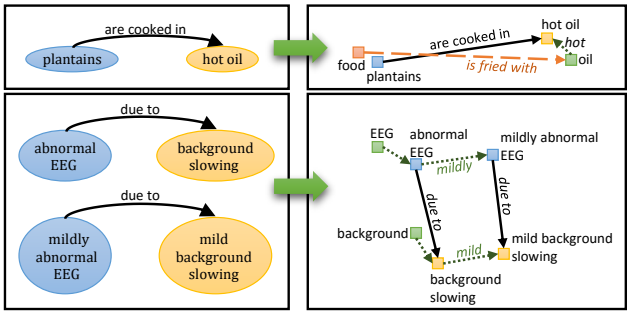

Figure 4: Examples of lexical relations encoded in the knowledge graphs and their representations in the embedding space for common-sense knowledge (top) and medical domain knowledge (bottom).

Figure 4 illustrates an example of common-sense knowledge and an example of medical domain knowledge in the knowledge graph (on the left) as well as in the embedded space (on the right). Each word is represented as a point in the embedded space, shown as a square. We capture the role of semantic composition and lexical relations as linear transformations in the embedded space, which are represented as arrows showing the point obtained after applying the transformation. For example, the top of Figure 4, shows how the relation $\langle$*plantains*, *are cooked in*, *hot oil*$\rangle$ is represented in the embedding space: a blue square for the word "plantains", a green square for the word "oil", and a yellow square for the phrase "hot oil". Figure 4 shows that applying the lexical relation "are cooked in" to the point "plantains" produces the same point as the phrase "hot oil." Likewise, it shows that the point for the phrase "hot oil" is obtained by applying the "hot" semantic composition vector to the word "oil". This example highlights the ability of the knowledge embeddings to infer new generalized relationships because the vector generated for the word "oil" in the phrase "hot oil" is generated such that it can be obtained by applying the "are cooked in" relationship to the word "plantains". Figure 4 also illustrates semantic composition for the domain of medicine: applying the transformation associated with the word "mild" (or "mildly") adjusts the "due to" relationship from "abnormal EEG" to "background slowing" such that when "due to" is applied to "mild background slowing", the phrase "mildly abnormal EEG" is produced instead of the more general concept "abnormal EEG". This demonstrates that the embedding procedure is able to account for the affects of individual words on the semantics of the discovered lexical relationships.

4.3.2. Accounting for Semantic Composition

Each vertex and each edge in our knowledge graph corresponds to a concept or relation which may be expressed with a multi-word sequence $w_1 \dots w_L$. To account for the contribution of each word on the semantics of the entire expression of concept or relation, we represented each word by a bag-of-words vector on which two composition functions are applied. The first composition function, denoted as $g(\bullet)$ defines uses the bag-of-words vector representations of each word expressing the concept to produce a new vector representing the entire concept. The second composition function,

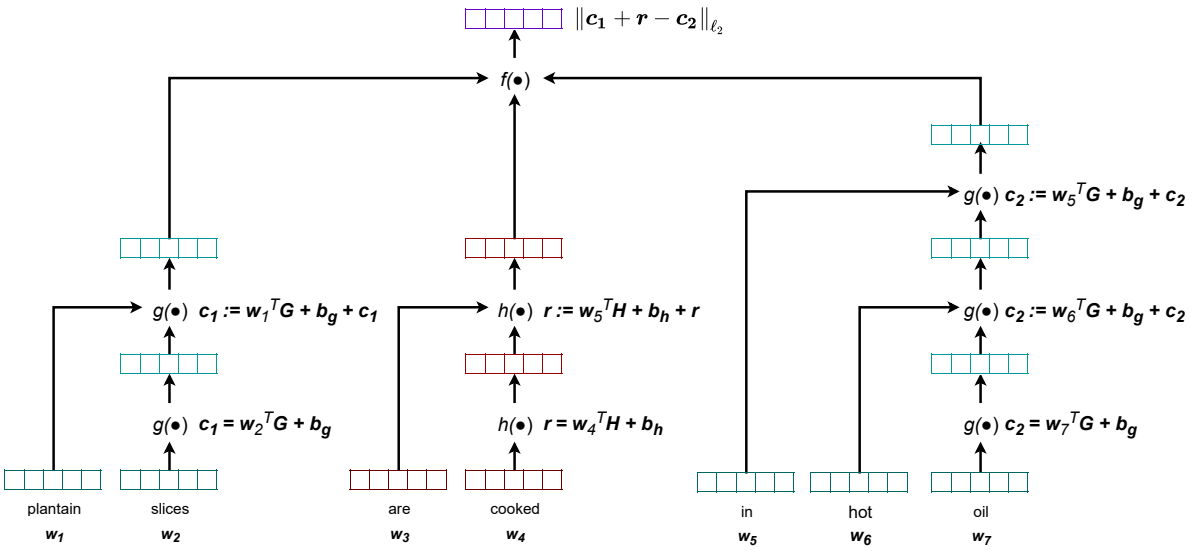

Figure 5: Neural architecture for computing the plausibility of a lexical reltion.

$h(\bullet)$, uses the bag-of-words vector representations expressing the lexicalized relation to generate a vector representation of the relation. Each of these composition functions was defined as a linear recurrence relation, which first maps the head word to a vector, and then recursively modifies that vector for each remaining word in the sequence. Formally, for each composition function we learned a composition matrix ($\boldsymbol{G}$ or $\boldsymbol{H}$) and a bias vector ($\boldsymbol{b_g}$ or $\boldsymbol{b_h}$) which captures the effect of each additional word on the vector for the word sequence. Formally, we define two semantic composition functions:

$$g(w_1 \dots w_L) = \begin{cases} \boldsymbol{w_L}^T \boldsymbol{G} + b_g, & \text{if } L = 1 \\ \boldsymbol{w_L}^T \boldsymbol{G} + b_g + g(w_1 \dots w_{L-1}), & \text{otherwise.} \end{cases} \quad (1)$$

and

$$h(w_1 \dots w_L) = \begin{cases} \boldsymbol{w_L}^T \boldsymbol{H} + b_h, & \text{if } L = 1 \\ \boldsymbol{w_L}^T \boldsymbol{H} + b_h + h(w_1 \dots w_{L-1}), & \text{otherwise.} \end{cases} \quad (2)$$

where $g(\bullet)$ recursively learns a vector for the each concept by repeated applying a linear transformation for each word in the concept using the learned semantic composition matrix $\boldsymbol{G}$, and the concept bias vector $\boldsymbol{b_g}$; and $h(\bullet)$ recursively learns a vector for each open-domain relation by repeatedly applying a linear transformation for each word in the relation using the learned semantic relation composition matrix $\boldsymbol{H}$, and the relation bias vector $\boldsymbol{b_h}$.

4.3.3. Evaluating the Plausibility of Knowledge

By representing the concepts and a lexical relations as vectors, we can measure the *plausibility* of any arbitrary lexical relation r between any possible pair of concepts, c_1 and c_2. Recall that we have represented each lexical relation as a geometric translation vector. This means, that for each concept c_1, the most likely concept to be related to it by lexical relation r should be embedded at the point $c_1 + r$. This allows us to measure the plausibility of any triple $\langle c_1, r, c_2 \rangle$ triple as the distance between the point in the embedded space most likely to be related to c_1 by r and the point corresponding by c_2, defined by:

$$f\left(\langle s_1, r, s_2 \rangle\right) = \|g(s_1) + h(r) - g(s_2)\|_{\ell_2} \quad (3)$$

Thus, equation 3 ensures the geometric property that the point associated with concept c_2 is obtained by applying the translation vector associated with r to the point for concept c_1, i.e., $g(c_2) \approx g(c_1) + h(r)$. Figure 5 illustrates the relationship between Equations 1, 2, and 3: the plausibility of a lexical relation depends on the semantic composition of each concept as well as the lexical relation.

4.3.4. Optimizing Knowledge Plausibility

By defining the plausibility of any structured relation triple $\langle c_1, r, c_2 \rangle$, we can learn the optimal values of the latent composition matrices $\boldsymbol{G}$ and $\boldsymbol{H}$, as well as the latent bias vectors $\boldsymbol{b_g}$, and $\boldsymbol{b_h}$. To do this, we try to find the values of these latent variables which maximize the plausibility of all *positive* edges in the knowledge graph and minimize the plausibility of *negative* edges which are not consistent with the knowledge graph. We construct so-called negative edges by randomly sampling edges from the knowledge graph and replacing the lexical relation with another, random lexical relation such that the new edge is not in the network. This allows optimal latent variables to be learned by comparing the *margin* (gap) in the geometric space between edges in the knowledge graph, and the negative edges which do not occur in the knowledge graph. Formally, we define the margin loss (Guo et al., 2015):

$$\mathcal{L} = \sum_{t^+ \in SN} \sum_{t^- \in SN^-} \max(0, \gamma + f(t^+) - f(t^-)) \quad (4)$$

where $t^+ = |s_1, r, s_2| \in SN$ refers to each triple in the knowledge graph, $t^- \in SN^-$ refers to artificially created *negative* triples, and γ, which indicates how much of a margin (or distance) should exist between positive triples encoded in the knowledge graph and the negative triples we randomly generated. As described in (Bordes et al., 2013; **?**), we applied stochastic gradient descent to solve this minimization problem.

4.3.5. Smoothing the Knowledge Embeddings

The embedded knowledge graph obtained by Equation 3 relies on distributional information in order to create an optimal embedding for each concept and each lexical relation the data. In order to ensure that the embeddings learned for the knowledge graphs are able to account for lexical variation, we introduce two regularization terms based on notions of lexical semantics. These terms are based on two assumptions: (1) vertices in the knowledge graph which have a high *functional similarity* (Turney, 2012) – that is, concepts which participate in many of the same relationships – should be located close to each other in the embedded space; and (2) edges in the knowledge graph which have a high *relational similarity* (Nakov and Hearst, 2008) – that is, lexical relations which often describe the same participants – should also be located close to each other in the embedded space. These assumptions allow the plausibility computed from our embedding space to account for lexical variation in concepts as well as lexical relations by ensuring that semantically similar concepts occupy geometrically close points (Guo et al., 2015). To do this, we define the following functional similarity matrix, $S_F(i,j)$ which returns the sum of the number of lexical relations in which both concepts i and j were participants. This allows us to measure the smoothness of similar vertices in the embedding space:

$$\mathcal{R}_1 = \frac{1}{2}\sum_{i=1}^{V}\sum_{j=1}^{V} \|\vec{v}_i - \vec{v}_j\|_{\ell_2}^2 \, S_F(i,j) \tag{5}$$

where V is the number of vertices in the knowledge graph. Equation 5 ensures that the distance between two concepts which participate in similar relations is small.

In order to ensure that lexical relations with high relational similarity have similar translation matrices, we construct a relational similarity matrix $S_R(i,j)$ which returns the total number of concepts for which lexical relations i and j are both edges in the knowledge graph. This allows us to account for the smoothness of similar edges in the embedding space:

$$\mathcal{R}_2 = \frac{1}{2}\sum_{i=1}^{E}\sum_{j=1}^{E} \|\vec{e}_i - \vec{e}_j\|_{\ell_2}^2 \, S_R(i,j) \tag{6}$$

where E is the number of edges in the knowledge graph. Equation 6 ensures that similar relations are embedded as similar transformations in the embedded space.

5. Experimental Results

Evaluating the quality of any common-sense knowledge base is known to be difficult (Singh et al., 2002). Consequently, we measured two aspects of our knowledge embedding: (1) the quality of the relations in the knowledge graph, and (2) the quality of any new inferred relations in the embedded space. When evaluating the quality of common-sense knowledge, we were inspired by previous efforts made to evaluate the OpenMind common-sense acquisition dataset (Singh et al., 2002). For each knowledge graph, we sampled 200 relations, and 200 *inferred* relations, obtained by sampling random vertices v and relations r and finding the vertices closest to the point obtained by $v + r$. Each of these common-sense relations were evaluated on a shifted 5 point Likert scale for the following desirable properties: (1) generality (such that -2 indicates a specific fact and +2 indicates a general statement about the world), (2) plausibility (such that -2 indicates a relationship which does not make sense in the world, and +2 indicates a relationship which makes complete sense), and (3) neutrality (such that -2 indicates highly-biased opinions and +2 indicates neutral sentiment). A total of three annotators were used, and obtained an inter-annotator agree of 84.3% (according to Cohen's kappa co-efficient). We evaluated both domains of common-sense knowledge: (1) general world knowledge and (2) medical knowledge. The average for generality was 0.8, and 1.2, indicating that while the knowledge from both sources was general, the medical knowledge was more-so. Likewise, the average rating for plausibility was 0.42 and 1.02, reflecting the fact that most medical relations stood alone as plausible facts while common-sense facts often rely on more context. This suggests that future work could benefit by incorporating coreference resolution. Finally, the average rating for neutrality was 0.05 and 1.85, suggesting that many of the relations encoded in the general dataset were biased by the opinions of the author, but that medical relationships were highly neutral.

6. Conclusions

In this paper, a framework for learning knowledge embeddings as representations of commonsense knowledge and domain-specific knowledge for medicine is presented. The novelty of the knowledge embeddings stems from the incorporation of semantic compositionality and of two regularization factors that accomplish the smoothing of the acquired knowledge - representing the new, un-articulated information discerned from vast text collections. To learn commonsense knowledge embeddings we have relied on a large set of blogs and books as well as the defining glosses from WordNet. To learn the knowledge embeddings for medicine, we have relied on two large set of electronic health records. We evaluated the quality of the common-sense knowledge obtained by the framework presented in this paper through manual review of randomly sampled relations encoded in the knowledge graph, as well as implicit relations inferred in the embedding space. The evaluation results for the two forms of knowledge that we have acquired are promising: the same knowledge acquisition methodology based on learning knowledge embeddings works well both for commonsense knowledge and for medical knowledge. Interestingly, the commonsense knowledge that we have acquired was evaluated as being less neutral than than the medical knowledge, because it often reflected the opinion of the knowledge writer. In addition, the medical knowledge that was acquired was evaluated as more plausible that the commonsense knowledge.

7. Acknowledgements

This work was supported by the National Human Genome Research Institute of the National Institutes of Health under award number 1U01HG008468. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

8. Bibliographical References

Akbik, A. and Michael, T. (2014). The weltmodell: A data-driven commonsense knowledge base. In *LREC*, pages 3272–3276.

Angeli, G., Premkumar, M. J., and Manning, C. D. (2015). Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL*, pages 26–31.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). *Dbpedia: A nucleus for a web of open data*. Springer.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, pages 2787–2795.

Burton, K., Java, A., and Soboroff, I. (2009). The icwsm 2009 spinn3r dataset. In *Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009), San Jose, CA*.

Cimino, J. J., Johnson, S. B., Peng, P., and Aguirre, A. (1993). From icd9-cm to mesh using the umls: a how-to guide. In *Proceedings of the annual symposium on computer application in medical care*, page 730. American Medical Informatics Association.

Goldberg, Y. and Orwant, J. (2013). A dataset of syntactic-ngrams over time from a very large corpus of english books. In *Second Joint Conference on Lexical and Computational Semantics (* SEM)*, volume 1, pages 241–247.

Goodwin, T. and Harabagiu, S. M. (2013). Automatic generation of a qualified medical knowledge graph and its usage for retrieving patient cohorts from electronic medical records. In *Semantic Computing (ICSC), 2013 IEEE Seventh International Conference on*, pages 363–370. IEEE.

Gordon, A. and Swanson, R. (2009). Identifying personal stories in millions of weblog entries. In *Third International Conference on Weblogs and Social Media, Data Challenge Workshop, San Jose, CA*.

Guo, S., Wang, Q., Wang, B., Wang, L., and Guo, L. (2015). Semantically smooth knowledge graph embedding. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pages 84–94.

Harati, A., Lopez, S., Obeid, I., Picone, J., Jacobson, M., and Tobochnik, S. (2014). The tuh eeg corpus: A big data resource for automated eeg interpretation. In *Signal Processing in Medicine and Biology Symposium (SPMB), 2014 IEEE*, pages 1–5. IEEE.

Lafourcade, M. (2007). Making people play for lexical acquisition with the jeuxdemots prototype. In *SNLP'07: 7th international symposium on natural language processing*, page 7.

Lenat, D. B. and Guha, R. V. (1989). *Building large knowledge-based systems; representation and inference in the Cyc project*. Addison-Wesley Longman Publishing Co., Inc.

Lowe, H. J. and Barnett, G. O. (1994). Understanding and using the medical subject headings (mesh) vocabulary to perform literature searches. *Jama*, 271(14):1103–1108.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., et al. (2011). Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182.

Miller, G. A. (1995). Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Nakov, P. and Hearst, M. A. (2008). Solving relational similarity problems using the web as a corpus. In *ACL*, pages 452–460.

Saeed, M., Villarroel, M., Reisner, A. T., Clifford, G., Lehman, L.-W., Moody, G., Heldt, T., Kyaw, T. H., Moody, B., and Mark, R. G. (2011). Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952.

Schank, R. C. and Abelson, R. P. (1995). Knowledge and memory: The real story. *Knowledge and memory: The real story. Advances in social cognition*, 8:1–85.

Schank, R. C. (1983). *Dynamic memory: A theory of reminding and learning in computers and people*. Cambridge University Press.

Singh, P., Lin, T., Mueller, E. T., Lim, G., Perkins, T., and Zhu, W. L. (2002). Open mind common sense: Knowledge acquisition from the general public. In *On the move to meaningful internet systems 2002: Coopis, doa, and odbase*, pages 1223–1237. Springer.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., et al. (2007). The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255.

Speer, R. and Havasi, C. (2013). Conceptnet 5: A large semantic network for relational knowledge. In *The People's Web Meets NLP*, pages 161–176. Springer.

Stearns, M., Price, C., Spackman, K., and Wang, A. (2001). Snomed clinical terms: overview of the development process and project status. In *Proceedings of the AMIA Symposium*, page 662. American Medical Informatics Association.

Turney, P. D. (2012). Domain and function: A dual-space model of semantic relations and compositions. *Journal of Artificial Intelligence Research*, pages 533–585.

Xin, R. S., Gonzalez, J. E., Franklin, M. J., and Stoica, I. (2013). Graphx: A resilient distributed graph system on spark. In *First International Workshop on Graph Data Management Experiences and Systems*, page 2. ACM.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., and Stoica, I. (2010). Spark: cluster computing with working sets. In *Proceedings of the 2nd USENIX conference on Hot topics in cloud computing*, volume 10, page 10.