Terminology Expansion with Prototype Embeddings: Extracting Symptoms of Urinary Tract Infection from Clinical Text

Mahbub Ul Alam[1], Aron Henriksson[1], Hideyuki Tanushi[2], Emil Thiman[2,3], Pontus Naucler[2,3] and Hercules Dalianis[1]

[1]*Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden*

[2]*Division of Infectious Disease, Department of Medicine, Karolinska Institutet, Stockholm, Sweden*

[3]*Department of Infectious Diseases, Karolinska University Hospital, Stockholm, Sweden*

Keywords: Natural Language Processing, Terminologies, Synonym Extraction, Word Embeddings, Clinical Text.

Abstract: Many natural language processing applications rely on the availability of domain-specific terminologies containing synonyms. To that end, semi-automatic methods for extracting additional synonyms of a given concept from corpora are useful, especially in low-resource domains and noisy genres such as clinical text, where non-standard language use and misspellings are prevalent. In this study, prototype embeddings based on seed words were used to create representations for (i) specific urinary tract infection (UTI) symptoms and (ii) UTI symptoms in general. Four word embedding methods and two phrase detection methods were evaluated using clinical data from Karolinska University Hospital. It is shown that prototype embeddings can effectively capture semantic information related to UTI symptoms. Using prototype embeddings for specific UTI symptoms led to the extraction of more symptom terms compared to using prototype embeddings for UTI symptoms in general. Overall, 142 additional UTI symptom terms were identified, yielding a more than 100% increment compared to the initial seed set. The mean average precision across all UTI symptoms was 0.51, and as high as 0.86 for one specific UTI symptom. This study provides an effective and cost-effective solution to terminology expansion with small amounts of labeled data.

SCIENCE AND TECHNOLOGY PUBLICATIONS

1 INTRODUCTION

In many applications of natural language processing (NLP), there is a need for ready access to domain-specific terminologies. However, for low-resource languages and domains, wide-coverage terminological resources tend to be scarce, and are often prohibitively expensive to create manually. In the context of noisy genres such as clinical text, where non-standard language use, creative shorthand and misspellings are prevalent (Dalianis, 2018), it is especially important to have access to domain-specific knowledge about the meaning of terms and their semantic relationships. To that end, semi-automatic and data-driven methods for extracting additional synonyms of a given concept from corpora are useful for expanding an existing but limited terminology. Such efforts are not only cost-efficient, but are also compelling due to their ability to capture real, domain-specific language use, including common spelling variants. This allows for the recall (sensitivity) of information extraction systems to be vastly improved.

Several different approaches to terminology expansion and synonym extraction have been proposed, including the use of lexico-syntactic patterns and graph-based models. More recent efforts have tended to leverage models of distributional semantics, especially in the form of word embeddings. These models are based on the distributional hypothesis (Harris, 1954), which states that words with similar distributions in a corpus – i.e. words that appear in similar contexts and co-occur with similar sets of words – often have similar meanings. They have become popular also in clinical NLP as data from electronic health records (EHRs) has become more readily accessible for research (Khattak et al., 2019). By creating vector-based representations of word meaning in semantic space, estimates of semantic similarity to other words in a corpus can be computed, forming the basis for many synonym extraction efforts in the clinical domain (Henriksson et al., 2014b; Zhang et al., 2017; Fan et al., 2019).

In this study, we explore and further investigate the notion of prototype embeddings for terminology

Alam, M., Henriksson, A., Tanushi, H., Thiman, E., Naucler, P. and Dalianis, H.
Terminology Expansion with Prototype Embeddings: Extracting Symptoms of Urinary Tract Infection from Clinical Text.
DOI: 10.5220/0010190200470057
In *Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2021) - Volume 5: HEALTHINF*, pages 47-57
ISBN: 978-989-758-490-9
Copyright © 2021 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

expansion. Prototype embeddings can be derived using any model of distributional semantics and are vector representations that aim to capture the meaning of higher-level concepts based on lexical instantiations of (some of) its members (Henriksson et al., 2014a). Prototype embeddings have been shown to be effective in generating semantic features that improve clinical named entity recognition systems; here, we build on the idea of prototype embeddings for expanding a terminology for urinary tract infection (UTI) symptoms by extracting candidate terms from clinical text corpora.

A UTI is an infection in any part of the urinary system, including kidneys, ureters, bladder and urethra. It is primarily caused by bacteria and is among the most common bacterial infections in the human body (Foxman, 2010). UTIs result in suffering and can also be lethal when they lead to sepsis (Herzog et al., 2014). Diagnosis of UTI is based on a combination of urinary symptoms and urine culture information (Rubin et al., 1992). There are a number of UTI symptoms, which can categorized as follows: painful urination (*dysuria*), frequent urination (*frequency*), constant urge of urination (*urgency*), tenderness in the lower abdomen (*suprapubic tenderness*), tenderness or pain elicited by percussion[1] from the kidney overlaying area in the back (*costovertebral angle*[2] *pain or tenderness*), as well as some other, less specific symptoms (*non-specific*) (ECDC, 2016; NHSN, 2017). Using only urine culture information for the diagnosis of UTI will lead to the overestimation of the incidence of UTI (Landers et al., 2010). As a result, the detection of UTI symptoms is critical for accurately identifying cases of UTI in EHRs.

While data-driven techniques that can be used to support terminology development are important for many domains, the specific motivation behind this study – from an application perspective – is to extract an extensive set of UTI symptom terms as they appear in real clinical text. The developed terminology of UTI symptoms is intended to be used for developing a system for automatically detecting UTIs based on structured and unstructured data from EHRs. The main contributions of this study are as follows:

- Two statistical phrase detection methods, with different thresholds, are explored to study the impact of the trade-off between the number and quality of the identified phrases on the downstream task of terminology expansion. Phrase detection is a necessary component in the data processing pipeline as many symptoms are multi-word expressions.
- Four word embedding methods are used for deriving prototype embeddings: *Word2Vec*, *Phrase2Vec*, *GloVe* and *FastText*. More importantly, we evaluate the use of prototype embeddings for terminology expansion and explore prototype embeddings at two levels of abstraction: (i) for specific UTI symptoms and (ii) for UTI symptoms in general.
- Two different corpora are used for training the prototype embeddings and we explore the trade-off between data volume and quality: one corpus is smaller but contains only positive UTI cases, whereas the other is larger but somewhat less relevant to the target domain.
- Using a small set of seed terms in the form of UTI symptoms, we are able to extract another 142 new terms for inclusion in the terminology using a data-driven and semi-automatic method based on prototype embeddings.

2 METHODS & MATERIALS

In this study, we investigate the use of prototype embeddings for the extraction of UTI symptom terms from clinical text. In order to create prototype embeddings for this task, we need: (i) to detect phrases in the unannotated corpus (in order to be able to capture symptoms that are not only expressed as single words but also as multiword expressions), and (ii) base embeddings from which to derive the prototype embeddings. We conduct a number of experiments with different underlying corpora, different methods for automatic detection of phrases, different methods for creating the base word embeddings, as well as prototype embeddings constructed at different levels of abstraction. Real-word clinical data is extracted from a major university hospital in Sweden. A domain expert annotates a portion of the data to create seed terms and also evaluates the candidate UTI symptom terms identified by the various models.

2.1 Methods

Below follows a description of the methods used in the study: (i) methods for phrase detection, (ii) methods for creating base word embeddings, and (iii) methods for creating prototype embeddings.

[1]'percussion' refers to the clinical examination process of tapping on the surface area of thorax or abdomen to determine the inner formation.

[2]'costovertebral angle' refers to the angle created by the vertebral column and the lower ending of the thorax.

2.1.1 Phrase Detection

For this task, phrase detection is necessary as symptom expressions can either be in the form of unigram words (e.g. *headache*) or multiword expressions (e.g. *sore throat*). Embeddings for terms of varying length therefore need to be created, and this is achieved by automatically identifying (and, in some cases, concatenating) phrases in the underlying corpus, which is later used for constructing the word embeddings.

Two common and simple data-driven phrase detection methods are used to this end. Both are based on the notion of identifying words that often co-occur together, but rarely in other contexts. The first one is presented in (Mikolov et al., 2013) and identifies phrases based on unigram and bigram counts according to the following scoring function:

$$score(w_i, w_j) = \frac{count(w_i w_j) - \delta}{count(w_i) \times count(w_j)},$$

where δ is a discounting coefficient that helps to avoid identifying too many phrases made up of very rare words. Bigrams that score above a certain set threshold are treated as phrases; this process is repeated in several passes over the data such that longer phrases than bigrams can also be identified. Here, this method is referred to as *IM* (for *iterative merging*). The second phrase detection method is based on the normalized (pointwise) mutual information among collocated words (Bouma, 2009). Here, we refer to this method as *nPMI*.

2.1.2 Word Embeddings

The distributional hypothesis (Harris, 1954) states that words that frequently co-occur in similar contexts tend to be semantically similar. Many methods have exploited this observation to automatically derive, from large text corpora, vector representations of word meaning. Word embeddings are lexical semantic representations in the form of dense, low-dimensional vectors in a continuous vector space, in which embeddings of semantically similar words are also expected to be in relatively close proximity in semantic (vector) space. Word embeddings can be derived using a number of different methods; as it has been shown that there is no single method that consistently outperforms others for all types of biomedical NLP tasks (Wang et al., 2018), we investigate the use of four common word embedding methods:

Word2Vec (Mikolov et al., 2013) derives, in an efficient manner, word embeddings using a shallow neural network that is trained to carry out a supervised learning task without the need for labeled data. There are two variants of the learning task: continuous bag of words (CBOW) and skip-gram. In CBOW, the task is to learn to predict the target word based on its context (i.e. the adjacent words in a fixed-size window), while, in the skip-gram model, the task is instead to predict the context based on the target word.

Phrase2Vec (Artetxe et al., 2018) is an extension of the former, designed to derive embeddings for phrases. This method requires one to provide a list of phrases separately, for which it learns phrase embeddings, along with regular unigram-based word embeddings.

GloVe (Pennington et al., 2014) combines global matrix factorization and local context window methods to derive word embeddings. The idea is to take into account the frequency of word co-occurrences in the entire corpus when deriving the word embeddings.

FastText (Bojanowski et al., 2017) treats words as a combination of n-gram characters. These n-gram characters can be mapped to dense vectors, and the overall aggregation of these lower-level embeddings can be used to represent a word or a phrase. This allows for deriving embeddings for unknown words and also requires less training data in comparison to the aforementioned methods.

2.1.3 Prototype Embeddings

Prototype embeddings (Henriksson et al., 2014a) are intended to capture the semantics of a (higher-level) concept or a group based on the embeddings of the members. A prototype embedding for a group can, for instance, simply be created through mean or median pooling of the members' embeddings. Here, we take the column-wise mean value of a set of embeddings to derive a prototype embedding. It has been shown that prototype embeddings can be used for creating semantic features that help to improve named entity recognition systems, while further improvements can be obtained by creating ensembles of prototype embeddings, where each member is derived from a model built with different underlying data and/or hyperparameters (Henriksson, 2015).

In this study, we investigate the use of prototype embeddings for a different task, namely terminology expansion, and, in particular, to extract terms – as they appear in real, clinical text – of UTI symptoms. We also investigate and compare the use of prototype embeddings constructed at different levels of abstraction: (i) one prototype embedding for each specific UTI symptom, and (ii) another, high-level prototype embedding for UTI symptoms in general. With the former, we aim to extract terms that express a specific UTI symptom and only use seed terms within that group to derive the prototype embedding. In the latter

case, we aim to extract any form of UTI symptom and use all seed terms to derive a single prototype embedding. We specifically investigate which abstraction level is the most productive for terminology expansion. Here, we refer to the former as *symptom-specific* and the latter as *symptom-general*.

2.2 Data

In this study, data in the form of text corpora are needed for constructing the word embeddings. In addition, the proposed method relies on access to seed terms for constructing prototype embeddings, both *symptom-specific* and *symptom-general*.

2.2.1 Corpora

The underlying corpora are extracted from a database of electronic health records from Karolinska University Hospital in Stockholm, Sweden. The data used in this study can be obtained (upon request) from the research infrastructure The Swedish Health Record Research Bank (Health Bank[3]), at Stockholm University (Dalianis et al., 2015). The infrastructure contains more than two million patient records from the years 2007-2014 obtained from Karolinska University Hospital.

We extracted clinical notes with the following inclusion criteria: (i) patients who are 18 years or older, (ii) admitted to the hospital between July, 2010 and March, 2013, and (iii) one urine culture taken during the hospitalization period. In total, there were 10,335 urine cultures found in 7,256 hospitalizations of 5,659 patients. A urine culture was considered positive if there was a significant growth (having more than or equal to 10^5 colony forming units per milliliter of urine) of no more than two pathogens. In total, there were 7,972 positive urine cultures found in 6,943 hospitalizations of 5,653 patients.

Table 1: Number of types and tokens in the two corpora.

Corpus	Types	Tokens
Case Group	156,695	13,475,706
Control Group	181,331	19,357,294

Two corpora are extracted for the experiments described later in section 2.3. One contains only clinical notes for hospitalizations that contain a positive urine culture, i.e. the *Case Group*. Another corpus is created with clinical notes for hospitalizations without a positive urine culture, i.e. the *Control Group*. The total number of types and tokens in the respective corpora are shown in Table 1. The corpora are preprocessed by removing punctuation marks and lowercasing all characters.

[3]http://dsv.su.se/healthbank

2.2.2 Seed Terms

In order to create a prototype embedding for a higher-level concept or group, access is needed to a sample of terms that represent members of that group. A physician and expert in infectious diseases, with extensive experience of treating patients with UTI, therefore manually annotated one month's (April, 2012) worth of data according to the aforementioned inclusion criteria. In total, 120 UTI symptom terms were annotated according to the six UTI symptoms mentioned in the introduction: *dysuria*, *frequency*, *urgency*, *suprapubic tenderness*, *costovertebral angle pain or tenderness*, and *non-specific*. In this annotation set, a total of 240 positive urine cultures were identified in 201 hospitalizations of 195 patients. The annotator marked the symptom terms with the exact form and spelling as found in the clinical text. Table 2 provides some examples of the annotated symptom terms. As can be seen, some symptom terms are misspelt (*trägnningar* should be *trängningar*); these need to be captured in order for the terminology to be effective for information extraction purposes. It is worth mentioning that the sixth UTI symptom (*non-specific*) was used to group the symptom terms which are not included in ECDC (European centre for disease prevention and control) or CDC (Centers for Disease Control and Prevention). We used it to group the terms which could still be relevant to detect UTI; for example, miktionsbesvär (micturition problems) could indicate some forms of disturbance related to micturition. The seed terms are also used for initial evaluation and hyper-parameter tuning, see section 2.3.4. Table 3 provides the number of manually annotated seed terms for each UTI symptom and their frequency in the two corpora.

2.3 Experimental Setup

In this paper, we investigate several research questions in the following sets of experiments:

2.3.1 Experiment 1: Underlying Data

One of the most fundamental aspects that affects the makeup of a word embedding space is the data which is used for training the model. Here, we investigate two aspects of the underlying data: (1) phrase detection and (2) data volume vs. quality.

Phrase detection is a necessary step in order to be able to identify UTI symptoms in the form of

Table 2: Examples of annotated UTI symptom terms.

UTI Symptom	Example Term	Translation
Dysuria	sveda	burning sensation
Frequency	kissar ofta	urinating often
Urgency	trägnningar	urgency (misspelt)
Suprapubic tenderness	ont i blåsa	bladder pain
Costovertebral angle pain or tenderness	flanksmärta	flank pain
Non-specific	miktionsbesvär	micturition problems

Table 3: Frequency of seed terms in the two corpora.

UTI Symptom	Case Group		Control Group	
	Types	Tokens	Types	Tokens
Dysuria	26	3,902	26	4,674
Frequency	9	337	9	395
Urgency	8	4,838	8	5,913
Suprapubic tenderness	14	49	14	55
Costo-vertebral angle pain / tenderness	35	1,254	35	1,495
Non-specific	28	1,701	28	2,067

multiword expressions. In data-driven approaches to phrase detection, there is a trade off between the number and quality of identified phrases. In addition to comparing the two data-driven phrase detection methods described in section 2.1.1, we explore the downstream impact of using a *small*, *medium*, or *large* list of automatically identified phrases. The phrase lists are generated using three different thresholds for each of the two phrase detection methods: 100 (small), 5 (medium), 1 (large) for *IM* and 0.57 (small), 0.34 (medium) and 0.23 (large) for *nPMI*. In order to ensure that the manually annotated UTI symptom terms are treated as phrases, they are concatenated using the underscore character ("_"). For example, all instances of "kissar ofta" (*urinating often*) are replaced by "kissar_ofta".

It is well-known that large corpora lead to higher-quality word embeddings, as it is necessary to have a large number of observations of language use – i.e. the contexts in which terms are used – in order to capture the variety and nuances of word meaning. However, the "quality" of word embeddings – here, defined according to their performance in the downstream task of terminology expansion – is also determined by the "quality" of the underlying data. In this context, we define data quality according to how specific the corpus is to the application domain of UTI. We investigate the use of two different underlying corpora: the *Case Group* corpus is relatively smaller but assumed to be of higher quality compared to the *Control Group* corpus, which is relatively larger but less specific to the application domain and hence assumed to be of lower quality. See Table 4 for the number of phrases identified with each setting and underlying corpus.

Table 4: The number of identified phrases in each corpus using different phrase lists (*Small*, *Medium*, *Large*) generated using two different phrase detection methods (*IM*, *nPMI*) and three different thresholds.

Phrase List	Case Group		Control Group	
	IM	*nPMI*	*IM*	*nPMI*
Small	7,780	7,145	11,149	10,233
Medium	29,918	28,626	41,896	40,728
Large	47,406	46,866	67,859	67,972

2.3.2 Experiment 2: Underlying Embeddings Method

Another important aspect that affects the makeup of a word embedding space is the method used for training the model. Different methods perform well in different domains and on different downstream tasks: here, we evaluate the following four word embedding methods to generate base models from which to derive prototype embeddings: *Word2Vec*, *Phrase2Vec*, *GloVe* and *FastText* (see section 2.1.2 for details).

The word embedding methods have many hyperparameters that need to be tuned. Instead of doing a grid search in some restricted hyperparameter space, points are chosen at random in order to more effectively search the space. For each word embedding method, 50 points are randomly selected, thus yielding 50 different models for each method. See Table 10 in the appendix section, which provides details concerning the hyperparameter space within which points are randomly sampled.

2.3.3 Experiment 3: Prototype Abstraction Level

One of the key research questions that we investigate in this study – building on previous work on the use of prototype embeddings – is on what level of abstraction prototype embeddings are best used for terminology expansion. In this study, we compare two

prototype abstraction levels: (1) at the specific UTI symptom level (*symptom-specific*), and (2) at the general UTI symptom level (*symptom-general*). All base word embedding models are used for deriving the best prototype embeddings within each abstraction level. The two levels are finally compared and evaluated for their ability to identify new UTI terms. The candidate terms produced by the prototype embedding models at each level are manually assessed by a domain expert, see section 2.3.4 for further details.

2.3.4 Evaluation

In this study, mean average precision (MAP) is used as the primary evaluation metric (Schütze et al., 2008). MAP is the simple average of average precision (AP) scores over all examples in a validation set. AP is a metric that describes to what extent relevant items are concentrated in the highest-ranked predictions. For each threshold level (k), AP can be calculated by first taking the difference between the *recall* at the current level in the ranked predictions and the *recall* at the previous threshold level ($k-1$), multiplied by the *precision* at that level (k) in the ranked prediction. The sum of the contributions at each level is the AP. *Precision* is the fraction of predictions that are relevant and correct, and *recall* is the fraction of all relevant values that are predicted.

For model selection, leave-one-out cross-validation is carried out. In this context, this entails that, in each iteration, all but one of the seed terms are used for deriving the prototype embedding; the ranking of the left-out seed term in the list of nearest neighbors – based on cosine similarity – is used for calculating the AP score. This process is repeated for all seed terms in order to estimate a MAP score for a given model. For *symptom-specific*, this process is carried out using seed terms for a specific UTI symptom, whereas for *symptom-general*, it is done using all seed terms. For *symptom-specific*, MAP scores are macro-averaged across the six UTI symptoms. For each abstraction level, the model with the highest macro-averaged MAP score is selected as the best model.

The best models within each level of abstraction and corpus are then compared and evaluated in the following manner. For both abstraction levels, all seed terms – for a specific UTI symptom or for all UTI symptoms, respectively – are used for constructing the prototype embeddings, i.e. there is no longer a need to leave out an instance. In total, 14 lists of candidate terms for inclusion in the terminology are generated. For each *symptom-specific* prototype embedding, the candidate list contains the terms corresponding to the 100 nearest neighbors. For each *symptom-general*, the candidate list contains the terms corresponding to the 600 nearest neighbors (6×100). A domain expert reviewed the union of the sets of candidate terms for relevance with respect to a certain UTI symptom. This allowed for counting the number of relevant UTI symptom terms that were extracted for each UTI symptom and abstraction level, as well as to calculate AP scores.

3 RESULTS

The first set of experiments were conducted using the initial set of seed terms for carrying out leave-one-out cross-validation. This allowed us to efficiently evaluate a number of potentially important factors in the creation of prototype embeddings for terminology expansion: (i) four different base embedding methods, (ii) two different phrase detection methods, each with three different thresholds controlling the number of phrases generated, and (iii) two different underlying corpora – one smaller but more relevant in scope, the other larger but less precise in terms of relevant scope. In Table 5 and 6, we present the results for *symptom-specific* prototype embeddings and *symptom-general* prototype embeddings, respectively. For each base embedding method and phrase detection method, we present results with the phrase list and corpus that yielded the best results.

For the *symptom-specific* prototype embeddings, as can be seen in Table 5, better results were obtained with *FastText* compared to the other base embedding methods, regardless of the phrase detection method used. The overall best result – a MAP score of 0.15 – was obtained with a *medium* phrase list obtained using the *IM* phrase detection method and the *Case Group* corpus. In these experiments, no clear difference was observed between the *Case Group* and *Control Group* corpora. With respect to the number of phrases identified, the results seem to speak in slight favor of using a small- or medium-sized phrase list.

For the *symptom-general* prototype embeddings, as can be seen in Table 6, the best results were again obtained using *FastText* as the base embedding method. Like in the case of *symptom-specific*, the overall best result – a MAP score of 0.14 – was obtained with a *medium* phrase list obtained using the *IM* phrase detection method and the *Case Group* corpus. Observations with respect to the choice of underlying corpus and phrase list are similar to the ones observed for *symptom-specific* prototype embeddings. The best-performing prototype embedding models at two different levels of abstraction (*symptom-specific* and *symptom-general*) and for

Table 5: Symptom-Specific prototype embeddings: macro-averaged MAP scores for different base embedding methods, phrase detection methods, the best phrase list and the best corpus. The highest scores for each phrase detection method are in bold.

Base Embedding	Phrase Detection	Phrase List	Corpus	MAP
Word2Vec	*IM*	*Medium*	*Control*	0.11
Phrase2Vec		*Large*	*Control*	0.10
GloVe		*Large*	*Case*	0.04
FastText		*Medium*	*Case*	**0.15**
Word2Vec	*nPMI*	*Medium*	*Case*	0.10
Phrase2Vec		*Large*	*Control*	0.11
GloVe		*Small*	*Case*	0.12
FastText		*Small*	*Control*	**0.12**

Table 6: Symptom-General prototype embeddings: macro-averaged MAP scores for different base embedding methods, phrase detection methods, the best phrase list and the best corpus. The highest scores for each phrase detection method are in bold.

Base Embedding	Phrase Detection	Phrase List	Corpus	MAP
Word2Vec	*IM*	*Medium*	*Control*	0.12
Phrase2Vec		*Large*	*Control*	0.10
GloVe		*Medium*	*Case*	0.07
FastText		*Medium*	*Case*	**0.14**
Word2Vec	*nPMI*	*Medium*	*Case*	0.12
Phrase2Vec		*Large*	*Control*	0.11
GloVe		*Small*	*Case*	0.13
FastText		*Small*	*Control*	**0.13**

two underlying corpora (*Case Group* and *Control Group*) were selected to participate in the final evaluation, wherein 100 candidate terms were extracted from each *symptom-specific* prototype embedding and 600 terms were extracted from each *symptom-general* prototype embedding. The candidate terms were reviewed by a domain expert for relevance and the results, in terms of AP scores, are shown in Table 7. All *symptom-specific* prototype embeddings perform well, with the exception of *suprapubic tenderness*. The *symptom-general* prototype embedding also performed well, but slightly worse compared to the macro-averaged MAP score for the *symptom-specific* prototype embeddings. Notably, using the *Case Group* corpus generally yielded better results with *symptom-specific* prototype embeddings (MAP: 0.51 vs. 0.48), whereas the *Control Group* corpus yielded better results with *symptom-general* prototype embeddings (MAP: 0.48 vs. 0.30).

Table 8 shows the number of extracted UTI terms (types) that were deemed relevant by the domain expert for each of the *symptom-specific* and *symptom-general* prototype embeddings, as well as the sum of their frequencies (tokens) in the two corpora. First, it

Table 7: Final evaluation: AP scores in the case and control corpora for each symptom-specific prototype embedding and the symptom-general prototype embedding.

Prototype Embedding	Case Group	Control Group
Dysuria	0.61	0.56
Frequency	0.64	0.47
Urgency	0.82	0.76
Suprapubic tenderness	0.00	0.06
Costovertebral angle pain or tenderness	0.86	0.83
Non-specific	0.13	0.24
Macro-averaged MAP	0.51	0.48
UTI Symptoms	0.30	0.48

should be noted that there were some terms that appeared in several of the candidate lists; the total number of unique candidate terms was 1,504. Of these, 142 terms were deemed relevant by the domain ex-

pert. The observant reader will notice that the sum of the types for the *symptom-specific* prototype embeddings is larger than 142 and this is because, in some cases, the domain expert classified a term as relevant for more than one specific UTI symptom. Nevertheless, more UTI symptom terms were extracted with the *symptom-specific* prototype embeddings than with the *symptom-general* counterparts (167 vs. 121). As expected, the terms are more frequent in the *Control Group* corpus, owing to its larger size.

Table 8: Frequency of the extracted and relevant UTI symptom terms in the two corpora.

Prototype Embedding	Case Group		Control Group	
	Types	**Tokens**	**Types**	**Tokens**
Dysuria	31	415	31	755
Frequency	43	367	43	527
Urgency	21	506	21	709
Suprapubic tenderness	27	98	27	131
Costo-vertebral angle pain / tenderness	9	510	9	759
Non-specific	36	765	36	1,081
UTI Symptoms	121	1,857	121	2,838

Table 9 provides an example of terms automatically extracted from a corpus of clinical text using a prototype embedding (and calculating its nearest neighbors), in this case for the UTI symptom *urgency*. As can be seen, many relevant terms are among the nearest neighbors of the prototype embedding. It is also notable that phrases of varying length are identified. There are also several misspellings, and the frequencies show that these are relatively common.

4 DISCUSSION

In this study, experiments were conducted concerning (i) the data and (ii) embedding methods used for constructing the semantic spaces, as well as (iii) the level of abstraction for the prototype embeddings. The results of these will be discussed below, in relation to the target application, namely terminology expansion and extracting UTI symptoms from clinical text.

The underlying data and embedding method used are naturally the two most important aspects that impact the structure of the resulting semantic space. Although these do perhaps not represent the primary focus of the paper, they were too important to ignore and we therefore studied their impact on the prototype embeddings that were, in turn, used for the downstream task of terminology expansion. Concerning the underlying data, this can be broken down into two parts: (i) phrase detection and (ii) corpus construction, in particular how the data is sampled and the trade-off between data volume vs. quality. In terms of the performance of the two phrase detection methods, there was little difference between them, with *IM* used in the best-performing models. When using statistical phrase detection methods, there is a clear trade-off between the number and quality of identified phrases; in this case, we could observe that using a large phrase list resulted in worse performance. As can be seen in Table 9, some of the identified phrases (e.g. *urinträngningar urinsticka*) are not phrases in a linguistic sense and, while deemed relevant by the domain expert, probably should not be included as terms in a terminology. While using linguistic information from a syntactic parser to generate phrases would likely yield better results, good syntactic parsers for low-resource languages and domains can be difficult to obtain, and the simpler methods used in this study generally produced satisfactory results. Moreover, using a vocabulary of standard phrases would be limiting since it would fail with the misspellings and the type of creative language use found in clinical text.

Regarding corpus construction, the results were mixed, making it difficult to draw any clear conclusions. However, in the final evaluation, it was observed that the *Control Group* corpus gave better results for *symptom-general* prototype embeddings and the *non-specific symptom-specific* prototype embedding, while the *Case Group* corpus gave better results for the other *symptom-specific* prototype embeddings (with the exception of *suprapubic tenderness*, which performed badly with both corpora). One possible explanation is that more data – even at the expense of being slightly less specific to the target domain – is helpful when the prototype embeddings are meant to capture concepts that are wider in scope, such as UTI symptoms in general or other, non-specific UTI symptoms. This would, however, need to be investigated further. A limitation with this experiment is also that the corpora are not all that different; in future work, it would be interesting to study this aspect in more detail and with greater differences between corpora, both in terms of volume and domain specificity.

Prototype embeddings are based on a notion that works with any vector-based model of distributional semantics. Our experiments showed that the choice

Table 9: Extracted symptom terms, along with English translations, for the prototype embedding for urgency. The ranks and the frequency in the Case Group corpus of relevant terms are shown. Misspelled terms are in bold.

Rank	Extracted Term	English Translation	Freq
1	trängningar vid miktion	urgency during micturation	15
2	besväras av täta trängningar	bothered by frequent urges	13
3	urinträngning	urinary incontinence	16
4	trängningarna	the urges	18
5	täta trängningar och sveda vid miktion	frequent urges and burning during micturition	11
6	täta urinträngningar	frequent urination	64
8	sveda och trängningar	burning and urges	30
9	täta trängningar till miktion	frequent urges for micturition	26
10	miktionsträngningar	micturition efforts	29
11	sveda vid miktion täta trängningar	burning during mictation frequent urges	16
12	miktionssveda och täta trängningar	micturition burns and frequent urges	13
13	upplever trängningar	experiencing urges	31
15	trängningar till vattenkastning	urge to urinate	11
16	trängningar till miktion	urges for micturition	46
18	täta miktionsträngningar	frequent micturition efforts	16
19	urinträngningar urinsticka	urinary incontinence urine stick	11
25	sveda eller trängningar	burning or urges	13
27	**trägningar**	urges	27
28	besvär med trängningar	discomfort with urges	11
37	form av trängningar	form of urges	12
38	trängningsbesvär	urgency	21
42	**täta trägningar**	frequent urges	15
63	**täta trängingar**	frequent urges	17

of base embedding method does have an impact on the downstream performance of the prototype embeddings. Among the ones included in this study, *FastText* consistently outperformed the others. There could be several explanations for this: one such explanation is that using subword embeddings allows it to generalize faster, and the corpora used in these experiments are both relatively small.

One of the key aspects we set out to investigate in this study, in addition to applying the notion of prototype embeddings to the task of terminology expansion, was to study if prototype embeddings could capture, not only synonymy, but something as wide in scope as UTI symptoms in general. While the performance was good with both *symptom-specific* and *symptom-general* prototype embeddings, with many new and relevant terms successfully identified, the former outperformed the latter in our experiments. In future work, it would be interesting to study this in more detail using a variety of concepts at different levels of abstraction, as well as to investigate the impact of the size and nature of the seed set used for deriving a prototype embedding. For example, using only one UTI symptom as seed terms, would it be possible to extract other types of UTI symptoms? One can also imagine more sophisticated ways of deriving prototype embeddings than mean pooling, even if simpler methods have certain advantages.

As can be seen in Table 7, the prototype embeddings indeed produced good results. Except for *suprapubic tenderness*, the performance was good in all cases, especially considering the frequency of these terms in the corpora. We looked for explanations for the poor performance of the *suprapubic tenderness* prototype embeddings and discovered that it was largely due to the low frequency of the associated symptom terms. The minimum frequency when creating word embeddings was set to ten and only two seed terms for this UTI symptom exceeded this threshold in the *Control Group* corpus, yielding an AP score of 0.06. In the *Case Group* corpus, only one seed term was present, resulting in an AP score of zero. In this case, it hence functioned like a regular word embedding to generate the candidate list, which, in turn, illustrates the advantage of prototype embeddings.

In future work, transfer learning will be explored, which involves fine-tuning a pre-trained model – trained with a large amount of data, not necessarily in-domain – to perform another task. In BERT (Devlin et al., 2018), multi-head attention is used to generate word embeddings. Due to its complexity and amount of data required, BERT-based models are typically used in transfer learning approaches, and we plan to explore this for terminology expansion. In future work, the terminology will be matched with the standard medical terminology available in Swedish, such as ICD-10 (international statistical classification of diseases and related health problems-10), Snomed

CT (systematized nomenclature of medicine – clinical terms), and MeSH (medical subject headings).

5 CONCLUSIONS

In this study, we investigated the use of prototype embeddings for terminology expansion, specifically for extracting symptoms of urinary tract infections from clinical text corpora. Four word embedding methods were used for deriving the higher-level prototype embeddings; it was observed that *FastText* yielded the best results. We also explored two statistical phrase detection methods and, while there was little difference between them, we also studied the trade-off between the number and quality of identified phrases and its impact on the downstream terminology expansion task. We also observed that using a somewhat smaller but high-quality, relevant corpus generally gave better results than using a larger yet less precise corpus; however, this seems to depend on the target concept's abstraction level. Indeed, two levels of abstraction were compared and contrasted: both yielded good results, but using prototype embeddings for specific symptoms overall outperformed the use of prototype embeddings for urinary tract infection symptoms in general. Ultimately, we were able to identify an additional 142 symptoms for inclusion in the terminology with very little manual effort required.

ACKNOWLEDGEMENTS

This research has been approved by the Regional Ethical Review Board in Stockholm under permission no. 2016/2309-32.

REFERENCES

Artetxe, M., Labaka, G., and Agirre, E. (2018). Unsupervised statistical machine translation. *arXiv preprint arXiv:1809.01272*.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Dalianis, H. (2018). *Clinical text mining: Secondary use of electronic patient records*. Springer, Open Access.

Dalianis, H., Henriksson, A., Kvist, M., Velupillai, S., and Weegar, R. (2015). Health bank-a workbench for data science applications in healthcare. In *CAiSE Industry Track*, pages 1–18.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

ECDC (2016). *Point prevalence survey of healthcare-associated infections and antimicrobial use in European acute care hospitals protocol version 5.3 : ECDC PPS 2016–2017*. ECDC, Stockholm.

Fan, Y., Pakhomov, S., McEwan, R., Zhao, W., Lindemann, E., and Zhang, R. (2019). Using word embeddings to expand terminology of dietary supplements on clinical notes. *JAMIA open*, 2(2):246–253.

Foxman, B. (2010). The epidemiology of urinary tract infection. *Nature Reviews Urology*, 7(12):653.

Harris, Z. S. (1954). Distributional structure. *Word*.

Henriksson, A. (2015). Learning multiple distributed prototypes of semantic categories for named entity recognition. *International journal of data mining and bioinformatics*, 13(4):395–411.

Henriksson, A., Dalianis, H., and Kowalski, S. (2014a). Generating features for named entity recognition by learning prototypes in semantic space: The case of de-identifying health records. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 450–457. IEEE.

Henriksson, A., Moen, H., Skeppstedt, M., Daudaravicius, V., and Duneld, M. (2014b). Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics*, 5(6).

Herzog, K., Dusel, J. E., Hugentobler, M., Beutin, L., Sägesser, G., Stephan, R., Hächler, H., and Nüesch-Inderbinen, M. (2014). Diarrheagenic enteroaggregative escherichia coli causing urinary tract infection and bacteremia leading to sepsis. *Infection*, 42(2):441–444.

Khattak, F. K., Jeblee, S., Pou-Prom, C., Abdalla, M., Meaney, C., and Rudzicz, F. (2019). A survey of word embeddings for clinical text. *Journal of Biomedical Informatics: X*, 4:100057.

Landers, T., Apte, M., Hyman, S., Furuya, Y., Glied, S., and Larson, E. (2010). A comparison of methods to detect urinary tract infections using electronic data. *The Joint Commission Journal on Quality and Patient Safety*, 36(9):411–417.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

NHSN (2017). *National Healthcare Safety Network (NHSN) Patient Safety Component Manual, Centers for Disease Control and Prevention; 2017*. NHSN, U.S. Department of Health & Human Services.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Rubin, R. H., Shapiro, E. D., Andriole, V. T., Davis, R. J., and Stamm, W. E. (1992). Evaluation of new anti-infective drugs for the treatment of urinary tract infection. *Clinical Infectious Diseases*, 15(Supplement_1):S216–S227.

Schütze, H., Manning, C. D., and Raghavan, P. (2008). Introduction to information retrieval. In *Proceedings of the international communication of association for computing machinery conference*, page 260.

Wang, Y., Liu, S., Afzal, N., Rastegar-Mojarad, M., Wang, L., Shen, F., Kingsbury, P., and Liu, H. (2018). A comparison of word embeddings for the biomedical natural language processing. *Journal of biomedical informatics*, 87:12–20.

Zhang, L., Li, J., and Wang, C. (2017). Automatic synonym extraction using word2vec and spectral clustering. In *2017 36th Chinese Control Conference (CCC)*, pages 5629–5632. IEEE.

APPENDIX

Table 10: Hyperparameter values for different word embedding methods.

Hyperparameter	**Values**
Corpus	Case, Control
Phrase detection method	IM, nPMI
Phrase list	Small, Medium, Large
Context window size	5, 10, 15
Vector dimension	50, 100
Iterations, GloVe	15, 20, 25, 30
Iterations, other methods	2, 5, 10
Hierarchical softmax value	1, 0
Skipgram value	1, 0
Negative value, Phrase2Vec	3, 5, 10
Negative value, other methods	5, 10, 15, 20
cbow_mean value for FastText	1, 0
Minimum term frequency	10
x max, GloVe	10
CBOW value, Phrase2Vec	0
min n, FastText	2
max n, FastText	10
Word ngrams, FastText	1

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS