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], Hercules Dalianis[1]

[1]**Department of Computer and Systems Sciences, Stockholm University, Stockholm, Sweden**
[2]Division of Infectious Disease, Department of Medicine, Karolinska Institute, Stockholm, Sweden
[3]Department of Infectious Diseases, Karolinska University Hospital, Stockholm, Sweden

mahbub@dsv.su.se
aronhen@dsv.su.se
hideyuki.tanushi@sll.se
emil.thiman@sll.se
pontus.naucler@ki.se
hercules@dsv.su.se

Stockholm University

UTI (Urinary Tract Infection), At A Glance

- An infection in any part of the urinary system, including kidneys, ureters, bladder and urethra
- Primarily caused by bacteria and is among the most common bacterial infections in the human
- Result in suffering and can also be lethal when they lead to sepsis
- Diagnosis of UTI is based on a combination of **urinary symptoms** and **urine culture information**
- Using only urine culture information for the diagnosis of UTI can lead to the overestimation of the incidence of UTI

Stockholm University

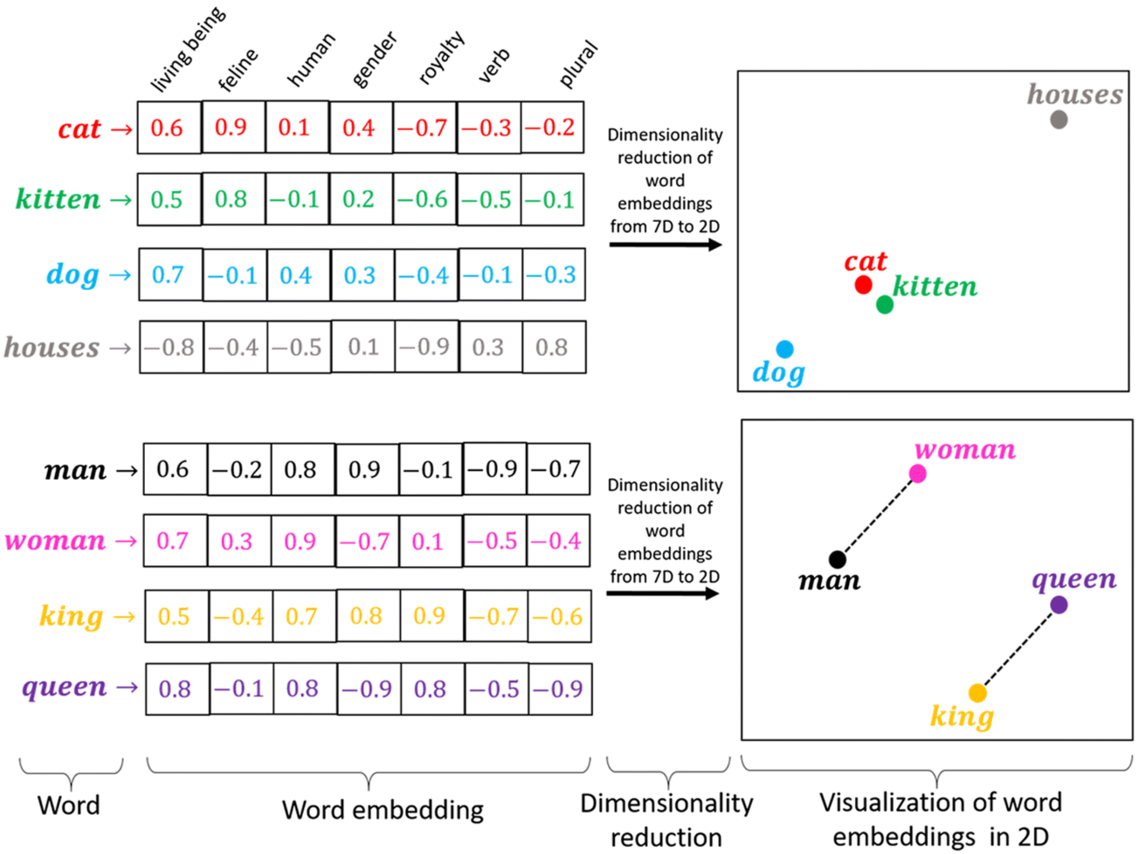
UTI, urinary symptoms and urine culture information

- painful urination (**dysuria**)
- frequent urination (**frequency**)
- constant urge of urination (**urgency**)
- tenderness in the lower abdomen (**suprapubic tenderness**)
- tenderness or pain elicited by percussion from the kidney overlaying area in the back (**costovertebral angle pain or tenderness**)
- other, less specific symptoms (**non-specific**)
- A **urine culture** can be considered **positive** if there is a significant growth
  - (having more than or equal to 10^5 colony forming units per milliliter of urine)

Goal

- Expanding a terminology for UTI symptoms by extracting candidate terms from a clinical text corpora using prototype embeddings
  - **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

Word Embedding



	living being	feline	human	gender	royalty	verb	plural
cat →	0.6	0.9	0.1	0.4	−0.7	−0.3	−0.2
kitten →	0.5	0.8	−0.1	0.2	−0.6	−0.5	−0.1
dog →	0.7	−0.1	0.4	0.3	−0.4	−0.1	−0.3
houses →	−0.8	−0.4	−0.5	0.1	−0.9	0.3	0.8
man →	0.6	−0.2	0.8	0.9	−0.1	−0.9	−0.7
woman →	0.7	0.3	0.9	−0.7	0.1	−0.5	−0.4
king →	0.5	−0.4	0.7	0.8	0.9	−0.7	−0.6
queen →	0.8	−0.1	0.8	−0.9	0.8	−0.5	−0.9

[1]

Words that appear in **similar contexts** and **co-occur with similar sets of words**, often have **similar meanings**.

Stockholm University

HEALTH BANK - Swedish Health Record Research Bank

- Unique research resource containing a large sets of electronic patient records
- Used in a number of research projects carried out by the Clinical Text Mining Group, Department of Computer and Systems Sciences, Stockholm University
- Contains data from over 512 clinical units from Karolinska University Hospital (2006–2014) over two million patients.
- Structured information contains, a serial number (de-identified) for each patient, age, gender, ICD-10 diagnosis codes, drugs, ab and blood values, admission and dicharge time, and date
- Unstructured data contains text written under different headings

Stockholm University

Data (1)

- Patients >= 18 years admitted to the hospital between July 2010 and March 2013
- One urine culture taken during the hospitalization period
  - 10,335 urine cultures found in 7,256 hospitalizations of 5,659 patients
  - 7,972 positive urine cultures found in 6,943 hospitalizations of 5,653 patients

- Two corpora are extracted
  - **Case Group**, contains only clinical notes for hospitalizations that contain a positive urine culture
    - 156,695 types, 13,475,706 tokens
  - **Control Group**, contains clinical notes for hospitalizations without a positive urine culture
    - 181,331 types, 19,35,294 tokens

Data (2)

- A physician and expert manually annotated one month's (April, 2012) worth of data to create seed terms
  - 120 UTI symptom terms were annotated according to the six UTI symptoms
  - A total of 240 positive urine cultures were identified in 201 hospitalizations of 195 patients

Example

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

Two Statistical Phrase Detection Methods

- **IM (iterative merging)**, identifies phrases based on unigram and bigram counts according to the following scoring function, where δ is a discounting coefficient that helps to avoid identifying too many phrases made up of very rare words,

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

- **nPMI (the normalized (pointwise mutual information)** among collocated words

Stockholm University

Four word embedding methods for deriving prototype embeddings

- **Word2Vec**, derives word embeddings using a shallow neural network
  - **Continuous bag of words (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)
  - **skip-gram**, the task is instead to predict the context based on the target word.
- **Phrase2Vec**, derives embeddings for phrases
  - Requires one to provide a list of phrases separately, for which it learns phrase embeddings

- **GloVe**, combines global matrix factorization and local context window methods to derive word embeddings
  - Takes into account the frequency of word co-occurrences in the entire corpus
- **FastText**, treats words as a combination of n-gram characters
  - n-gram characters can be mapped to dense vectors
  - The overall aggregation of these lower-level embeddings can be used to represent a word or a phrase
  - Allows for deriving embeddings for unknown words
  - Requires less training data in comparison

Experiments

- **Experiment 1: Underlying Data**
  - Phrase detection
  - Data volume vs. quality

- **Experiment 2: Underlying Embeddings Method**
  - Evaluating the four word embedding methods to generate base models from which to derive prototype embeddings

- **Experiment 3: Prototype Abstraction Level**
  - At the specific UTI symptom level (symptom-specific)
  - 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

Evaluation Metrics

- **Mean average precision (MAP)**, simple average of average precision (AP) scores over all examples in a validation set
- **Average precision (AP)**, 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**, the fraction of predictions that are relevant
- **Recall**, the fraction of all relevant values that are predicted

Stockholm University

Best Model Evaluation Criteria (1)

- Leave-one-out cross-validation is carried out
  - 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
  - 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

Best Model Evaluation Criteria (2)

- 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
  - 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

Identified Phrases

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

Stockholm University

Symptom-Specific Prototype Embeddings

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**

Symptom-General Prototype Embeddings

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**

Final Evaluation

Candidate terms were re-viewed by a domain expert for relevance and the results, in terms of AP scores

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

Stockholm University

Frequency of Seed Terms & Extracted Relevant terms In The Two Corpora

Seed Terms

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

Extracted Relevant Terms

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

Example, Extracted Symptom Terms

- 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

Discussion (1)

- There was little difference between the two phrase detection methods, with IM used in the best-performing models
- Using a large phrase list resulted in worse performance
- Control Group corpus gave better results for symptom-general prototype embeddings and the non-specific symptom-specific prototype embedding
- Case Group corpus gave better results for the other symptom-specific prototype embeddings

Stockholm University

Discussion (2)

- The choice of base embedding method does have an impact on the downstream performance of the prototype embeddings
  - FastText consistently outperformed the others

- Symptom-specific prototype embeddings outperformed the symptom-general prototype embeddings
- Ultimately, we were able to identify an additional 142 symptoms for inclusion in the terminology with very little manual effort required
  - A more than 100% increment compared to the initial seed set

Questions?

Image Source

[1] https://medium.com/@hari4om/word-embedding-d816f643140

Stockholm University

Backup Slides

Hyperparameter values

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