Vocabulary mapping with word embeddings
AMLD 2020

Pekka Tiikkainen

Motivation

- Planning of epidemiology studies often involves mapping terms between vocabularies.
- Challenge: at Roche Drug Safety, medical events are often defined with MedDRA but RWD databases use ICD9-CM and ICD10-CM to encode data.
- Mapping tables exist but they are incomplete.
- The objective of this work was to improve mapping coverage using word embeddings.

Examples of known mappings

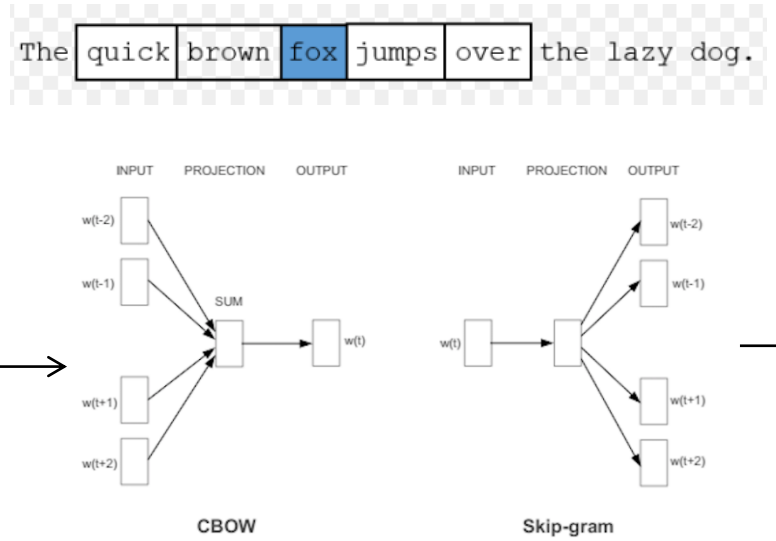
MedDRA preferred term	ICD term	Mapping type
Amnesia	Memory loss (ICD9-CM)	Exact
Pulse absent	Other specified symptoms and signs involving the circulatory and respiratory systems (ICD10-CM)	Approximate
Non-small cell lung cancer	?	Missing

Word embeddings in a nutshell

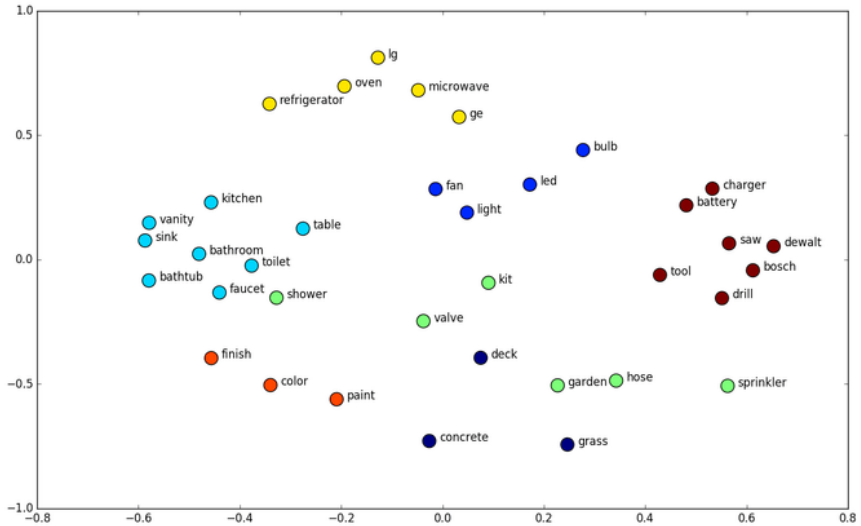
Word embeddings represent words as numeric vectors.
The more similar the context the words appear in, the more similar the vectors.

Various algorithms exist. The diagram below represents word2vec which is probably the most popular one.

(Large) text corpus,
e.g. All wikipedia articles, all Pubmed abstracts...

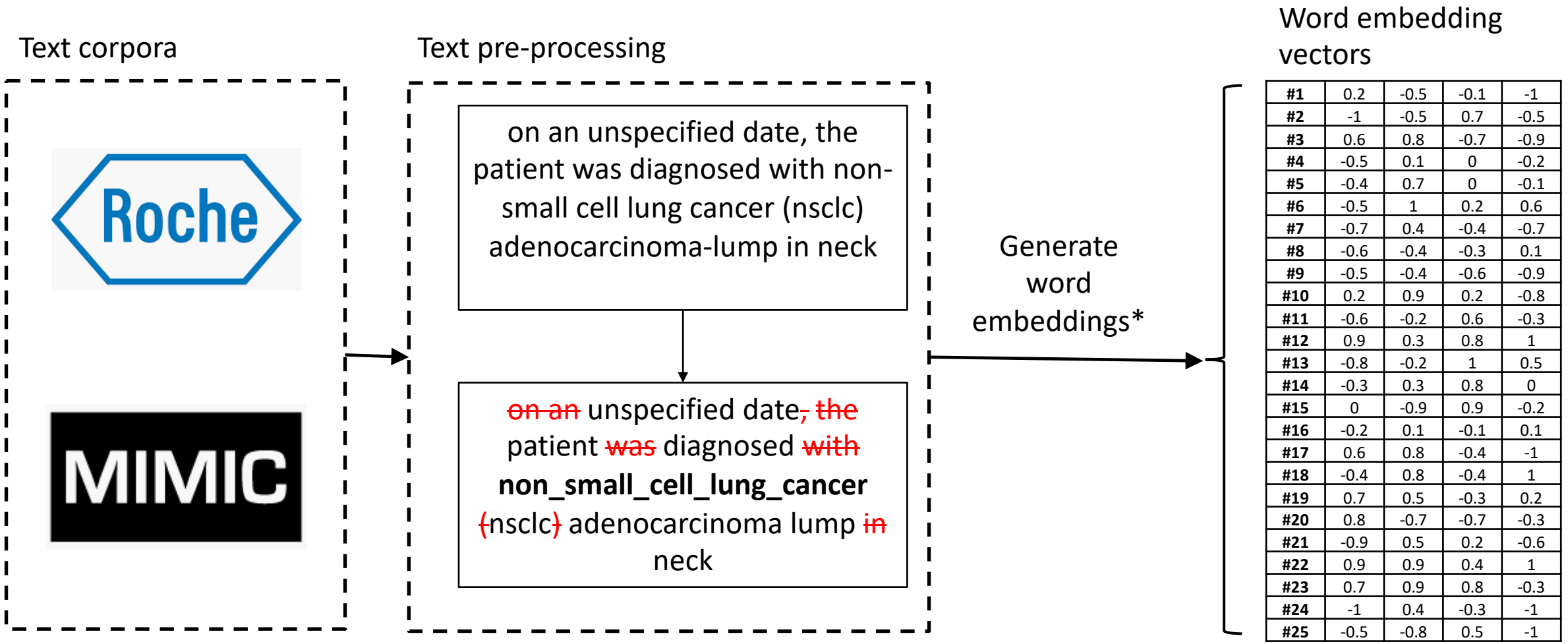


Using either CBOW or Skip-gram, train a neural net. In the end, extract the projection layer (numeric vector for each word).



Use the vectors to cluster words, calculate distances, as input to machine learning models etc.

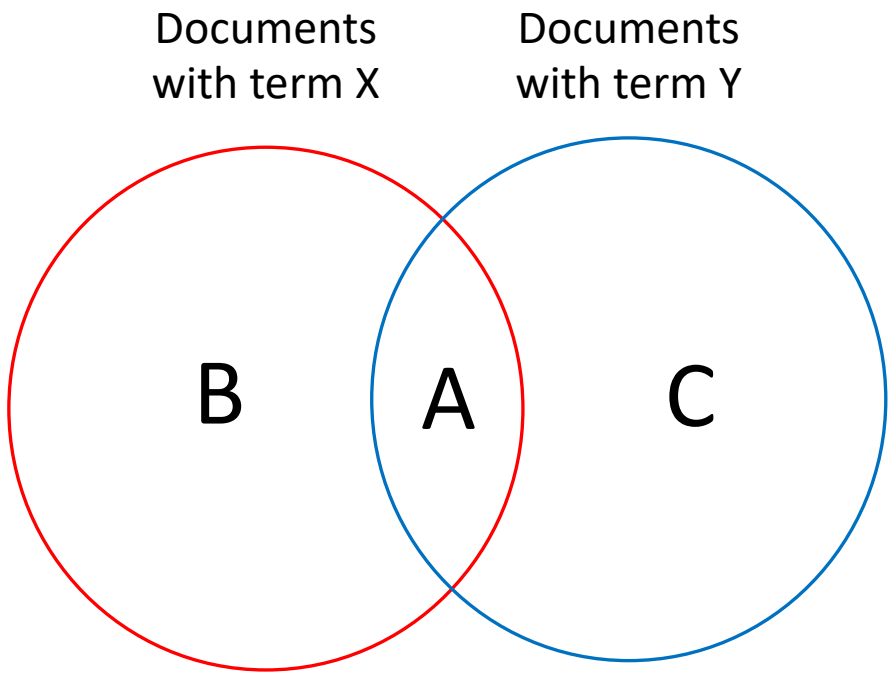
Training word embeddings for MedDRA and ICD9/10



* algorithms: word2vec, GloVe and Fasttext with various hyperparameter combinations

Baseline

- To be worth the extra effort, the more complex word embedding methods should outperform the baseline.
- A simple baseline distance metric based on term co-occurence was calculated.



$$\text{Baseline distance} = 1 - \frac{|A|}{|B|+|C|-|A|}$$

Example 1

Closest ICD10-CM terms to the MedDRA term «non-small cell lung cancer».

Most are terms related to lung cancer.

Interestingly, also pancreatic cancer terms show up. Potential explanation is that pancreatic cancer often metastasises in the lung.

Distance	Code	Term name
0,432	C34.82	malignant neoplasm of overlapping sites of left bronchus and lung
0,440	J91.0	malignant pleural effusion
0,441	C34.81	malignant neoplasm of overlapping sites of right bronchus and lung
0,446	C34	malignant neoplasm of bronchus and lung
0,463	C34.80	malignant neoplasm of overlapping sites of unspecified bronchus and lung
0,472	C34.92	malignant neoplasm of unspecified part of left bronchus or lung
0,473	C33	malignant neoplasm of trachea
0,477	C34.9	malignant neoplasm of unspecified part of bronchus or lung
0,478	C34.91	malignant neoplasm of unspecified part of right bronchus or lung
0,490	C34.0	malignant neoplasm of main bronchus
0,500	C34.02	malignant neoplasm of left main bronchus
0,501	C34.90	malignant neoplasm of unspecified part of unspecified bronchus or lung
0,507	C34.01	malignant neoplasm of right main bronchus
0,512	C34.00	malignant neoplasm of unspecified main bronchus
0,519	C25.2	malignant neoplasm of tail of pancreas
0,525	C38.4	malignant neoplasm of pleura
0,533	C25	malignant neoplasm of pancreas
0,535	C25.1	malignant neoplasm of body of pancreas
0,543	C77.1	secondary and unspecified malignant neoplasm of intrathoracic lymph nodes
0,550	C43.59	malignant melanoma of other part of trunk
0,550	C45	mesothelioma
0,553	R04.2	hemoptysis
0,560	C25.0	malignant neoplasm of head of pancreas
0,561	C78.0	secondary malignant neoplasm of lung
0,568	C78.00	secondary malignant neoplasm of unspecified lung
0,569	C43	malignant melanoma of skin
0,569	C78.2	secondary malignant neoplasm of pleura
0,570	C67	malignant neoplasm of bladder
0,574	C78.02	secondary malignant neoplasm of left lung
0,575	C18.4	malignant neoplasm of transverse colon
0,577	C77	secondary and unspecified malignant neoplasm of lymph nodes
0,583	C71.3	malignant neoplasm of parietal lobe
0,583	C61	malignant neoplasm of prostate
0,584	C56	malignant neoplasm of ovary

Embedding algorithm: Fasttext
Corpus: Roche safety DB

Example 2

Closest ICD10-CM terms to the MedDRA term «pulse absent».

As one would expect, cardiac events are prominent on the list.

Distance	Code	Term
0,363	I49.01	ventricular fibrillation
0,373	I46	cardiac arrest
0,424	I47.2	ventricular tachycardia
0,425	R09.2	respiratory arrest
0,434	R23.0	cyanosis
0,461	I49.02	ventricular flutter
0,475	I47	paroxysmal tachycardia
0,496	R57.0	cardiogenic shock
0,514	R06.4	hyperventilation
0,516	I95	hypotension
0,532	R06.0	dyspnea
0,534	H57.04	mydriasis
0,537	R40.2	coma
0,556	R06.1	stridor
0,559	I47.1	supraventricular tachycardia
0,559	R41.2	retrograde amnesia
0,565	T71	asphyxiation
0,566	G93.82	brain death
0,572	J98.1	pulmonary collapse
0,576	I50.84	end stage heart failure
0,577	I31.4	cardiac tamponade
0,577	R40.1	stupor
0,578	I50.81	right heart failure
0,580	J96.0	acute respiratory failure
0,580	G82.5	quadriplegia
0,580	J81	pulmonary edema
0,580	R25	abnormal involuntary movements
0,584	E87.2	acidosis
0,584	R23.1	pallor
0,584	R06.03	acute respiratory distress
0,586	I50.82	biventricular heart failure
0,589	R09.02	hypoxemia
0,589	R00.2	palpitations
0,589	R47.81	slurred speech
0,590	Z66	do not resuscitate
0,590	I21	acute myocardial infarction

Embedding algorithm: Fasttext
Corpus: Roche safety DB

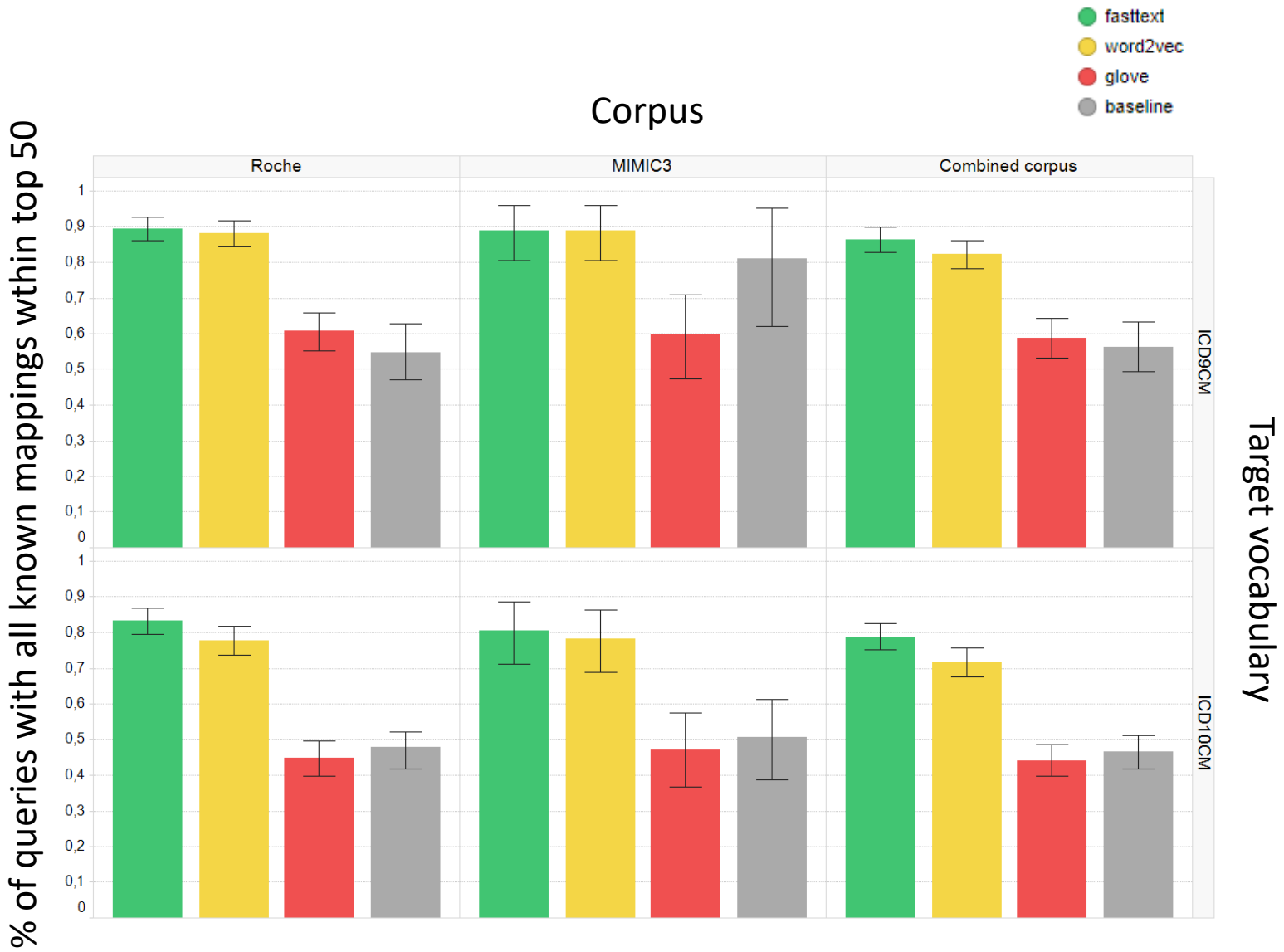
Performance

Using a table of known mappings for 524 MedDRA terms, different combinations of corpora, algorithms and parameters were evaluated.

For each MedDRA term, the 50 nearest ICD terms were listed. The bars on the rights give the proportion of terms which had all their known mappings in the top 50.

Key findings:

- Fasttext and word2vec perform best and outperform the baseline.
- Performance appears to be better when mapping to ICD9CM.
- Choice of corpora makes little difference.



Wrap-up

- A novel approach for medical terminology term mapping.
- Meant to complements existing approaches
- Validation with known mappings shows that word embeddings are able to find medically relevant mappings
  - However, the performance may be lower for MedDRA terms for which no mappings exist yet
- Goal is to build a company-wide tool for term mapping.
  - The tool would suggest a list of matching terms,
    final decision to be made by the medical expert using the tool

Doing now what patients need next