## Lecture 2: Electronic Medical Records

CSCI6XXX/CHE6XXX/CSCI4XXX (CSCI6093)

Finlay Maguire (finlay.maguire@dal.ca)

#### Learning Outcomes

- Describe electronic medical/health record systems and the types of data they typically contain
- Distinguish structured, semi-structured, unstructured text data
- Describe approaches to searching text
- Outline key steps in preparing text for analysis
- Explain the general concept of learnt word embeddings
- Explain how embeddings can be tuned/customised
- Identify differences between named entity recognition, parts of speech tagging, and dependency parsing

- Not covered: fuzzy search and text indexing

### What is an Electronic {Medical,Health} Record?

#### EMR are digital patient charts

- Repository of patient information over time
- Prone to fragmentation between primary / hospital care
- Ideally contains all of a given patient's details on:
  - Every encounter with health professionals (e.g., admitted to hospital)
  - Details and results of diagnostic testing and vitals (e.g., blood test, urine cultures etc.)
  - Diagnostic/therapeutic orders (e.g., Nil per os/NPO,
  - Procedures performed (e.g., appendectomy, PET-scan)
  - Medical note (e.g., primary physician, consult information)
  - Medication (e.g., antibiotics)

	Year														Year	
	Month 1	Month 2		Month 3	Mo	nth 4	Month 5	Month	6	Month 7	Month 8	Month 9	Month 10	Month 11	Month	12
Medication					Ø				_	+0		00000				
Notes	O	O		0	0					0	000					
Diagnoses	0	0	0	0	0		0	0	0	0	D O			$\bigcirc$		
Procedures	0	0		0	0		0	0	C	0	0 0	00		0		
Orders	0	Ø		0	0		0	0	0	0	0 00 0	$\bigcirc \phi \bigcirc \phi \bigcirc \phi$		0		
Labs & Flowsheets	s O	0	0	0	0	0	0	0	C	0	0 0	00000	0000			
Encounters	00	000		$\phi$ $\otimes$	000	00	00	Ø	00 00	$\mathbf{O} \circ \mathbf{O}$	000	000 000000		$\odot$		

#### Patient Timeline

#### 10.1038/s41746-018-0029-1

#### EMR data varies in structure

Patient Timeline



#### EMRs have a surprisingly long history



https://mavaberlerner.medium.com/a-brief-history-of-ehrs-c51a2125a247



#### EMRs are common and increasing in use in Canada

- Use of EMRs is common and increasing in Canada (2017-2021: 82% to 87%)
- Primary care is main users (93%) compared to specialists (80%)
- Atlantic regions have lowest use rate (74%)
- Main features used:
  - Ordering diagonstic tests / accessing results
  - Prescription system with automatic warning of adverse drug interactions
  - Communication of discharge/consultation notes
- Not used:
  - Clinical decision support (<27%)
  - Appointment scheduling (<21%)
  - AI/Machine Learning (<2%)

MAIN EMR SOFTV	VARE	SPEC	IALTY	*		R	EGIO	N			SE	TTING	;
NAIDED AMONG THOSE WITH EI	MR, n=1,771)	GP (960)	Spec. (829)	<b>ATL</b> (132)	QC (116)	<b>ON</b> (584)	<b>MB</b> (94)	<b>SK</b> (68)	<b>AB</b> (325)	BC (443)	Comm. (813)	Hosp. (394)	Both (564)
Accuro EMR	17%	16%	18%	11%	-	14%	66%	51%	12%	17%	17%	6%	23%
Med Access	11%	18%	4%	19%	-	-	1%	39%	17%	19%	14%	2%	14%
PS Suite EMR	10%	15%	3%	4%	2%	27%	-	-	5%	-	12%	1%	12%
OSCAR	7%	10%	3%	1%	-	9%	-	-	-	18%	9%	1%	8%
Epic EMR	7%	1%	13%	-	-	16%	-	-	8%		1%	22%	4%
Wolf EMR	5%	8%	1%	-	-	-	-	-	15%	8%	9%		5%
Meditech 📕	4%	1%	7%	13%	-	7%	-	-	-	4%	-	11%	4%
Profile by Intrahealth 📕	3%	5%	1%	15%	-	-	-	-		9%	4%	1%	5%
Cerner 📕	3%	1%	6%	2%	-	5%	-	-		6%	-	10%	2%
Connect Care	3%	1%	5%	-	-	-	-	-	11%	1%	1%	4%	3%
Telus (Unspecified)	3%	3%	2%	3%	1%	5%	-	-	2%	1%	4%	1%	2%
Healthquest EMR	2%	3%	1%	-	-	3 <del>-</del> 53	-		11%		5%	1%	1%
PLEXIA EMR	2%	1%	4%	-	-	-	-	-		9%	2%	2%	3%
Sunrise Clinical Manager	2%	-	5%	1%	-	-	-	14%	7%		1%	6%	1%
Medesync 📕	2%	3%	2%	-	27%	-	-	-	-	-	3%	-	2%

NOTE: Included mentions reported by 2%+ of sample answering

2021 National Survey of Canadian Physicians

#### One Patient One Record Experience

- Mix of Meditech, SHARE, individual silos
- One Patient One Record
  - 2014 December: Strategy Approval
  - 2015 July:
    - NSHA CIO appointed to lead project
    - RFP solutions hired to monitor procurement process to lead OPOR
  - 2016 March: Meeting moratorium
  - 2016 May/December: Gevity/Allscripts NSHA meetings
  - 2017 January: Request for Supplier Qualifications
  - 2017 February: Submissions from 4 big firms (**Epic**, Allscripts, Cerner, **Meditech**) and 2 small ones (Evident, Harris Healthcare Group)
  - 2017 June: Allscripts and Cerner named as only finalists based on 50 page RFSQ
  - 2017 August: Evident complaint
  - ....
  - ???
  - ...
  - ~2022: One Patient One Experience

### Reality: fragmented EMRs requiring difficult linkage

- Interoperability is a competitive disadvantage
- Standarised format and vocabulary for EMR data:
   Observational Medical
   Outcomes Partnership
   (OMOP) common data
   model
- OHDSI tooling
- Fast healthcare interoperability resources (FHIR)



https://ohdsi.github.io/TheBookOfOhdsi/images/CommonDataModel/cdmDiagram.png

#### EMR datasets: MIMIC-.\* / Canadian Primary Care Sentinel Surveillance Network / STARR



10.1038/sdata.2016.35

https://informatics.bmj.com/content/27/3/e100161

#### EMR allow us to ask complex questions



## What kind of data is in an EMR?

#### Many types of data in EMRs

- Discrete Physiological Parameters e.g., blood test metric measures
- Diagnostic Imaging Data e.g., MRI image data
- Physiological Sensor Data e.g., EKG/EEG signal data
- Ordinal scale assessments e.g., frailty index

#### Text:

- <u>Structured text</u> e.g., CPT/ICD-10 codes (*V89.2XXA, S06.0*)
- <u>Semi-structured Text</u> e.g., {"Symptoms": "Head pain, dizziness, emesis", "Cause": "Car crash", "Diagnosis": "Likely concussion"....}
- <u>Unstructured Text</u>: "Patient was involved in a car crash and presented to the ER with pain, vomiting, and mild dizziness. Most likely they are concussed but should follow up with a head CT to confirm no other brain injuries"

CAT SCANS								
ABDOMEN		ICD-10	DESCRIPTION					
Abdomen w/o contrast 74150	ormations and	d Chromosomal	Abnormalities (Including Down's Syndrome)					
Abdomen w/ contrast 74160	EC	Q64.79	Other congenital malformations of bladder and urethra					
Abdomen w/o & w/ contrast 74170		Q90.9	Down syndrome, unspecified					
	10S	Q05.4	Unspecified spina bifida with hydrocephalus					
CHEST/THORAX		Q05.8	Sacral spina bifida without hydrocephalus					
Chest/Thorax w/o contrast	nitourinary System Diseases (Including Incontinence and UTI)							
Chest/Thorax w/ contrast	OS	N17.9	Acute kidney failure, unspecified					
Chest/Thorax w/o & w/ contrast								

#### Medicine loves unstructured text

- Unstructured text is and will forever remain the primary form of communication in medical clinical settings.
- Highly flexible, efficient, and expressive across a range of communication contexts for medicine.
- Mainstay of charts, notes, consults, discharge summaries, procedure/operative logs.

S or $U$ (standarmached with of $T_{\rm FR}$ (subplace toler-up) of $T_0$ (subplace matrix $R_{\rm e}$ (subplace)	Date:	<del></del>	
adaution with lower personals	HCP IDW:	·	OFFICE VISITS PADE A CONTE
		1.100	Vice (1, , , , , , , , , , , , , , , , , , ,
Physical Assessment	*D	*ND	All the second s
Temperature pm	10 CONSC 1		VEDRESS 13 Jack Avenue Avenue (Norma) (Norma)
Palse pm	122		000008108 at all to 000 19(2-05-05 Ade 97
Respiration pm	10° - 3		Contract E and an entral in the second secon
Blood Prossure per			ALCINE DESCRIPTION OF A
Health pm	1.0		ADATE/VITAL SKIND ADATO ADOTO ADATO ADOTO ADATO ADOTO ADATO ADOTO ADATO ADOTO ADATO ADOTO ADOT
Weightom			I PANS I TO THE
Chest sounds of visit			1 Sale and 34 p.0 (P) OC - Coursely + transmission to the child
07 Set of evaluation	10	and the set	
Asthena Control:	D.	ND	HET. as hell the of wome whereas & chief
Cruch where or chest foltness (cdd/wh5		1100	
scolling to the state of the state of the state			Deverative they train & I year, The sympler shared
Wate Scient (CLW)	1.2	-	and the product of the second second second second
Physical activity broded due to adhread		-	- Mentine Method and the second contract, a manual day
Nach Related with statics	10	-	Wet Site The El almerica by adher exceptionary
Nanda Bullanor J. Andra	-	-	
Even admitting (boarded admit #B obsit Wells in ("hints) stress last stati-			another more higher thank executly chicking for the 14
exactles and the product and the rest of the start of the product and the	1.12		The ste accounts of the second s
School-work absurce since last visit	13		a program ban for filling a state of the state
Spirametry	D	ND	is the not well y she has when y to
FEV Lpre			dust & largered constrain alles another that the
FEV I post			
FEV1 %charge	1.0		Bit bussend small but office the proper
PEFpte			de instruction de la company
PEFpost	- E1		Care when Jains your Southan Sis Start ag
PEP % charge	11 12	2420	I we are ever mult light and an along in the dark
Review	0	ND	The state of the second st
Definition of Aefirms	22 ACC 2		Test pro test tot
Action Plan - (ACP)			] Wey Searry phase with Englis - third 15 080
Action Plan - (methal)	12		Barry Barry Co Chest 4172 HEISE . 3 AUG. MAD.
Med. Admin. Technique	1		the second of the second second second second
Warning signs	1.00		SHIT I HAD Y Y Y
Trigger factors	23 5		personal of the spin text por allo the then Game
Environmental control	22 2		The series with the series of
Coping strategies			Sprending with a set of the set of 14 and the the the the
Medications			I CHARLEN DE COMPANY IN THE COMPANY INTERCE.
Current			Mod Saludenet Bit & year Sale due Ante he Sign of 78 Bay
			Bid Brenning for piter of har Jacons on is a result
			that to be ashe with deduction .
Frescribed			The Can Have done to the interest highest manage
			SICCP I LOOK GYA CA. THE THE SALE AND A MALEY AND
			the bad & where areas & start back
Nonitor potential side effects	D	ND	File Father, dry of the production of your shit wo
height/ostroporosis, etc.	24 8728 3	1 2 3 3 1	mayber the SAL. I was shirt
Referrals	D	ND	Devid He & Smooth DETER, DRS & Dur Chuldren
Authma Education Program	10 10 CC		historial in a supportive house he ld draw allow the
Asthma Support Group	100		and a solution of a solution of the solution of the
Specialist	- S2		
the first second s		~	

#### Unstructured text is challenging

- English language especially has many synonymous and highly flexible grammatical structure.
- Medical english has even more synonymous:
  - Bilateral salpingectomy
  - Salpinoectomy
  - Fallopian Transection
  - Fallopian Tubectomy
  - Fallopian Tubal Ligation
  - Tubal ligation
  - Tubal sterilisation
  - Tubal
  - CPT58600
- Now add typos and transcription errors!
- Difficult to search
- Difficult to summarize
- Difficult to analyze



## So how would we do something like this?

### Natural Language Processing!

- NLP is any computer-based method that handles/augments/transforms natural language so that it can be represented for computation.
- Approximate synonyms: text mining, text processing, computational linguistics
- Example problem:

1. "Find every medical note in the EMR related to the spine"

2. Identify key search terms e.g., "back", "spine", "vertebra", "lumbar", "neck", "cervical", "thoracic", "sacrum", "coccyx" (expertise, ontology/vocabularly)

3. Search for EMR for these terms



## Let's start simple: searching text

#### Searching for exact matches: Ctrl-F

Many exact match algorithms with varied properties (typically ctrl-F will mix and match them in a context-dependent way).

- Scan over all text and look for things that exactly match your query
- Make things more efficient: Boyer-Moore/Knuth-Morris-Pratt/Rabin-Karp etc.

*u* doesn't occur in *P*, so skip next two alignments

P: word
T: There would have been a time for such a word
word skip!
word skip!
word skip!
word

#### More flexible searches for keywords: Regular Expressions

- Need to find "spine" and "spinal" = *spin(a|e)/*?
- Can also be used to capture words/before after: \w+\sspin(a|e)/?\s\w+
- Builds on lots of well-developed CS theory

 You have a problem, you use regex, you now have 2 problems

Character	Description	Example
[]	A set of characters	"[a-m]"
١	Signals a special sequence (can also be used to escape special characters)	"\d"
	Any character (except newline character)	"heo"
^	Starts with	"^hello"
\$	Ends with	"world\$"
*	Zero or more occurrences	"aix*"
+	One or more occurrences	"aix+"
{}	Exactly the specified number of occurrences	"al{2}"
Ι	Either or	"falls stays"
()	Capture and group	

#### Regular expressions can get very complicated!

RCF5322 Email validation regex:

 $(?:[a-z0-9!\#\%\&''+/=?^_`{}]*-]+(?:\[a-z0-9!\#\%\&''+/=?^_`{}]*-]+)*|"(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21\x23-\x5b\x5d-\x7f]|\[\x01-\x09\x0b\x0c\x0e-\x7f])*")@(?:(?)$ 

:[a-z0-9](?:[a-z0-9-]\*[a-z0-9])?\.)+[a-z0-9](?:[a-z0-9-]\*[a-z0-9])?\\[(?:(?:(2(5[

 $0-5] | [0-4] [ 0-9] | 1 [ 0-9] [ 0-9] | [1-9]? [ 0-9] ) | .) \{3\} (?: (2(5[0-5] | [0-4] [ 0-9] ) | 1 [ 0-9] [ 0-9] ] | 1 [ 0-9] [ 0-9] | 1 [ 0-9] ] | 1 [ 0-9] [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9] | 1 [ 0-9$ 

-9]|[1-9]?[0-9])|[a-z0-9-]\*[a-z0-9]:(?:[\x01-\x08\x0b\x0c\x0e-\x1f\x21-\x5a\x53-\

x7f]|\\[\x01-\x09\x0b\x0c\x0e-\x7f])+)\])

# Can we make the text easier to search instead?

#### Most NLP methods start with text normalisation

- 1. Tokenisation
- 2. Normalising word formats
- 3. Segmenting sentences

### Splitting text into words: Segmentation/Tokenizing

- Breaking text into individual units (letters/morpheme/words/sentences/paragraphs) can make it much easier to handle.
- Process is known as tokenisation (a subset of segmentation)
- Units you break into are known as tokens:

"Indication of significant spinal contusions."

-> "Indication" "of" "significant" "spinal" "contusions."

- Easy approach:
  - split on spaces
  - Has to be fast (finite state automata)
- Challenges: punctuation can matter (e.g., 01/02/22), not all languages use spaces, may want to treat multiword expressions (MWE) as tokens e.g., "New York", "bilateral salpingo-oophorectomy", "ice box"/"ice-box"/"icebox"

#### Simplifying language: word normalisation

- "Ph.D.", "PhD", "phd" probably shouldn't be counted differently
- Case folding: collapse everything to lowercase (although case can often be informative: "US" vs "us")
- Lemmatization: identifying words with common root (lemma) e.g., "operation" and "operations" -> "operation"; "am", "are", "is" -> "be"
  - "Surgeon is performing surgical procedures" -> "Surgeon be perform surgical procedure"
- Requires morphological parsing splitting <u>stems</u> (central morpheme) from <u>affixes (modifying/adidtional meaning)</u>
- Lemmatization is difficult : alternative = stemming
  - Remove final affixes e.g., remove "-ing, -s, -ational, -sses"
  - "This was not the correct operation" -> "Thi wa not the correct operat"

#### Splitting sentences: Sentence Segmentation

- Sentences are delineated on punctuation: ".", "?", "!"
- Often we want to segment phrases/clauses, more challenging:
  - "Patient presented to ER with pain/confusion, most likely as a sequelae of a head injury" ->
    - ["Patient", "presented", "to" "ER", "with", "pain", "confusion"]
    - ["most", "likely", "as", "a", "sequelae", "of", "a", "head", "injury"]

#### Hash-based text search

- Can find exact matches very efficiently
- Tokenize/lemmatise/normalise words in each note, then hash:
  - Note 1: ["spine", "car", "head"] -> [a11, a92, a53]
  - Note 2: ["car", "tree", "CT"] -> [a92, a57, a99]
- Hash query words "back", "spine", "vertebra", "lumbar", "neck", "cervical", "thoracic", "sacrum", "coccyx":
  - a55, a11, ...
- See if query hashes are present in note hash sets
  - a55 in Note1 = No, a55 in Note2 = No
  - a11 in Note1 = Yes, a11 in Note 2 = No

#### Can use these to create manual rules

if finding is in ("pneumothorax";
"hydropneumothorax")

and certainty-modifier is not in

("no"; "rule out"; "cannot evaluate")

and **status-modifier** is not in ("resolved")

then

conclude true;

endif;



https://doi.org/10.7326/0003-4819-122-9-199505010-00007

# What if we don't know the query terms in advance?

#### Identify frequently used terms



- Find highest TF-IDF terms
- Filter them manually for new search terms
- Apply prior search approaches (or any of the fuzzy matching approaches)
- Among other
   unsupervised approaches
   (e.g., following material)

### Can we automate more complex procedures?







#### **Template Filling**

Tumor Temp	olate
Reference:	Mass
Location:	Medial right lobe
Size:	7.2 cm

## Identifying tokens referring to things: Named Entity Recognition

- Identify specific categories of entities e.g., places, times, anatomy
- Lots of pre-trained approaches/vocabularies
- Text classification problem => requires some way to encode text to a numerical vector

revealed a 7.2 - cm	n mass in the medial right
lobe without evidenc	se of ductal dilation
	Name entity recognition Predetermined atomic concert are identified
revealed a 7.2 - cn	n <mark>mass</mark> in the <mark>medial</mark> right
lobe without evidence	ce of ductal dilation,

#### Encoding text as a vectors

One-Hot encoding

Very large vectors with even moderate vocabularly size

Very sparse vectors (lots of 0s)



The quick brown fox jumped over the brown dog

							$\checkmark$						
	cat	the	quick	brown	fox	jumped	over	dog	bird	flew	. 1	kangaro	o hous
!	0	1	0	0	0	0	0	0	0	0		0	0
	0	0	1	0	0	0	0	0	0	0		0	0
	0	0	0	1	0	0	0	0	0	0		0	0
	0	0	0	0	1	0	0	0	0	0		0	0
	0	0	0	0	0	1	0	0	0	0		0	0
	0	0	0	0	0	0	1	0	0	0		0	0
	0	1	0	0	0	0	0	0	0	0		0	0
	0	0	0	1	0	0	0	0	0	0		0	0
	0	0	0	0	0	0	0	1	0	0		0	0

#### **Document Vectorization**



Sum over columns for each note to get a vector representation of the document instead (TF-IDF is a normalisation of this representation)

#### Reducing the dimensionality of these vectors

- Standard dimensionality reduction methods struggle
- Text has semantic AND syntactic aspects/similarity
- We want to find a lower dimensional <u>embedding</u> that captures these aspects
- "You shall know a word by the company it keeps (Firth, J. R. 1957:11)"
- "The meaning of a word is its use in the language" (Wittgenstein)
- Can we use the **CONTEXT** of a given word to find a meaningful vector representation?
- Answer:
  - Byte-Pair Encoding
  - N-grams
  - Learned Word Embeddings



#### Learned word embeddings



Word2Vec -> GloVe -> ELMo -> BERT -> ERNIE -> GPT-3/Megatron/T5

#### Learnt embeddings are powerful



#### Custom embeddings and fine-tuning

- Same approach can be used beyond just words:
- Med2Vec
- EHR2Vec
- BioALBERT
- Corpora used to create embedding may not be a good fit for specialised text (i.e., EMRs aren't representative of the internet at large... we hope).
- Repeat training on your data but initialise with pre-trained weights



# With embeddings we can build/use models for more complex problems

## Train classifier on labelled medical text (e.g., ontology) = Named Entity Recognition



#### **Template Filling**

Tumor Temp	olate
Reference:	Mass
Location:	Medial right lobe
Size:	7.2 cm

## Train document classifier on EMR notes labelled by organ => Text classification



#### **Template Filling**

Tumor Template						
Reference:	Mass					
Location:	Medial right lobe					
Size:	7.2 cm					

## Use classifier trained to identify parts of speech and their relations (previously HMMs)



#### Overview

- Describe electronic medical/health record systems and the types of data they typically contain
- Distinguish structured, semi-structured, unstructured text data
- Describe approaches to searching text
- Outline key steps in preparing text for analysis
- Explain the general concept of learnt word embeddings
- Explain how embeddings can be tuned/customised
- Identify differences between named entity recognition, parts of speech tagging, and dependency parsing

- Not covered: fuzzy search and text indexing