Clinical Natural Language Processing

Modern Clinical Data Science

May 13, 2021

Mount Sinai

Outline of the Class

1. Introduction
2. Defining the Task and Approach
3. Clinical Information Extraction
4. Embeddings and Pretraining
5. Text Classification
6. Additional Topics: Weak Supervision and NLI

Part I
Introduction

What is NLP?

(From Wikipedia) A subfield of linguistics, computer science, information engineering, and artificial intelligence concerned with… how to program computers to process and analyze large amounts of natural language data.

- Automatic speech recognition
- CCG
- Common sense
- Constituency parsing
- Coreference resolution
- Dependency parsing
- Dialogue
- Domain adaptation
- Entity linking
- Grammatical error correction
- Information extraction
- Language modeling
- Lexical normalization
- Machine translation
- Missing elements
- Multi-task learning
- Multi-modal
- Named entity recognition
- Natural language inference
- Part-of-speech tagging
- Question answering
- Relation prediction
- Relationship extraction
- Semantic textual similarity
- Semantic parsing
- Semantic role labeling
- Sentiment analysis
- Shallow syntax
- Simplification
- Intent Detection and Slot Filling
- Stance detection
- Summarization
- Taxonomy learning
- Temporal processing
- Text classification
- Word sense disambiguation

https://github.com/sebastianruder/NLP-progress

Why “NLP” clinical text?

Most of the information we want to use as features or outcomes in clinical predictive models is found in the text of clinical notes (e.g., laboratory, pathology, or progress notes).

- Treatment goals and outcomes (success or failure of treatments, criteria for success, decisions about subsequent treatments)
- Interpretations of radiology and pathology images and laboratory test results
- Social determinants of health (e.g., social connection/isolation, housing issues, mentions of financial resource strain)
- Symptoms, symptom changes, and their interpretation
- Past medical history and family history
- Primary and secondary complaints
- Psychiatric evaluations and records of therapy sessions
- Patients’ emotional disposition, mood, and interactions with health providers
- Detailed descriptions of procedures (e.g., labor and delivery, heart catheterization, imaging studies, surgeries)
- Adherence to treatment plans (e.g., medications, physical therapy, procedures)
- Allergies, side effects, and other adverse events
- Results of physical examination (e.g., review of systems and interpretation of findings)
- Patients’ reasons for seeing a health provider
- Discharge summaries and follow-up plans

Why “NLP” clinical text? (Example from i2b2)

note-88.txt

```
995247940 | ABOHS | 01328259 | | 983241 | 8/25/1998 12:00:00 AM | CEREBRO VASCULAR
ACCIDENT | Signed | DIS | Admission Date: 4/10/1998 Report Status: Signed

Discharge Date: 1/19/1998
PRINCIPAL DIAGNOSIS: LEFT SUBCORTICAL STROKE.
PROBLEM LIST: 1 )
3 ) HYPERTENSION.
5 ) STATUS POST T
CERVICAL DYSPLASI
SITES OF ARTHRITI
8 ) VITILIGO. 9 )
HISTORY OF PRESEN
presentation , wh
return to bed , t
and right arm clu
all day long. She
She has remained
prior to admissio
admission. This e
paralysis ( she h
waited for it to
greater than 24 h
had not taken ins
intake secondary
difficulty swallo
REVIEW OF SYSTEMS: She had no fevers , chills , vomiting or loss
of consciousness. No seizure activity. No
head or neck trauma. No vertigo or tenderness. No recent change
in medications. No chest pain or shortness of breath , no
photophobia or neck stiffness.
PAST MEDICAL HIST
goiter , hyperten
trait , status po
cervical dysplasi
complaints , just
hemorrhoids.
MEDICATIONS: Hydr
Norvasc 5 milligr
once a day , Prem
insulin 12 units
14 units of regul
also been taking
pain. She is also
once a day.
SOCIAL HISTORY: T
a widow. No clear
use. She has two
mobility.
FAMILY HISTORY: Familial intermittent paralytic syndrome.
PHYSICAL EXAMINATION: Temperature was 97.1 , heart rate 110 ,
respirations 20 , blood pressure 130/70 ,
oxygen saturation 100% on room air. Glucose is 388. Mental status
examination: The patient was alert and oriented times three
following three step command across the midline. No apraxia.
Positive dysarthria. Repetition intact. Memory 3/3 immediately
and 3/3 at 5 minutes. Cranial nerves: Pupils 3-2 bilaterally.
Right central facial paralysis. No pallor or droop. Decreased
sensation to light touch , pinprick and temperature on the right.
Tears well bilaterally. She cannot shrug her right shoulder.
Motor and tone flaccid on the right. Strength: Right hemiplegia
( can only wiggle toes ) , left strength 4-5/5 ( poor effort ). Deep
tendon reflexes are 2+ at the biceps , 1+ at the knees , 0-1 at the
ankle with an upgoing right toe. Sensation: Decreased sensation
to light touch , pinprick and vibration. Temperature on the right
including trunk. Gait not assessed. HEENT: Atraumatic. Neck
supple. Lungs clear to auscultation bilaterally. Cardiovascular:
Regular rate and rhythm. Abdomen: Soft and nontender. Positive
bowel sounds. Extremities: No edema.
EKG: Sinus tachycardia at 104 per report when compared with EKG
of January , no significant change. Chest x-ray: No
evidence of active cardiopulmonary disease. CT scan of the head on
8/17/98 showed low density area present in the posterior limb of
the left intracapsular capsule consistent with an infarct.
Otherwise , no significant abnormality was seen. Mild age
appropriate involutional changes were present.
```

Unique Features of Biomedical/Clinical Text

- Specialized vocabulary / abbreviations
- Sentence fragments (esp. clinical)
- Long, complex sentences with multiple clauses (esp. scientific)
- High degree of assumed knowledge

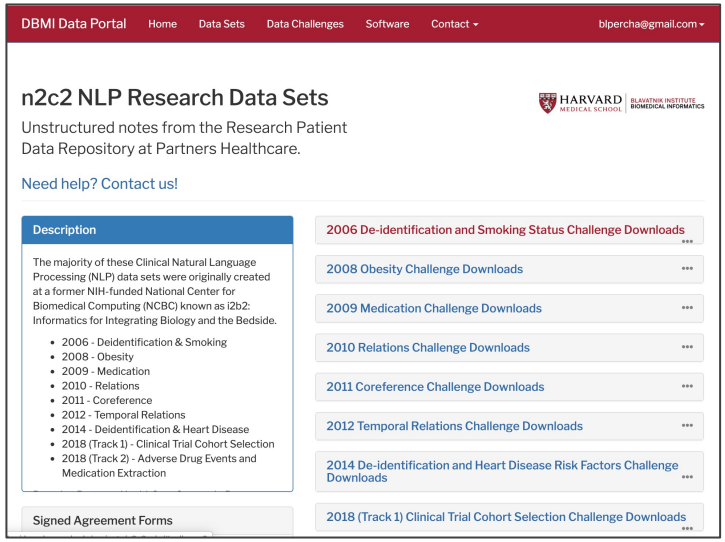
It’s important to remember that

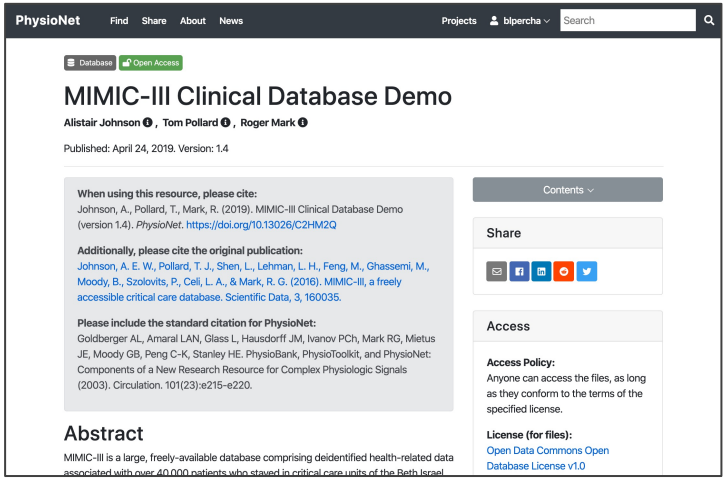
NLP is not “solved”.

Consider Google search, Siri, Alexa, etc.

Biomedical and Clinical Corpora

- MIMIC III
- i2b2 (now n2c2)
- PubMed
- PubMed Central
- BioCreative
- GENIA corpus
- Others (see Wikipedia "Biomedical Text Mining" article)





Mount Sinai Data Warehouse (MSDW)
(slide courtesy of Patricia Kovatch)

9.3 million Number of **patient records** in MSDW	**4.5 million** Number of **patients with clinical data** in MSDW collected since 2003	**80 million** Number of **patient visits** recorded, including: inpatient, outpatient and ED	**107 million** Number of **ICD-9 and ICD-10 coded diagnoses** in MSDW collected since 2003
	370 million Number of **test results**, including: lab, radiology and pathology	**189 million** Number of **medication records**, including prescription and med admins	**45 million** Number of **clinical documentations** like progress notes, discharge summaries & operative reports

- **Self service, de-identified access through the Cohort Query Tool (CQT), TriNetX, i2b2 Atlas/OMOP**
- **Identified data available by request (fee)**
- **Integrated with BioME and the Image Research Warehouse (>200,000 images)**
- **Creating a database of unstructured physician notes for Natural Language Processing (NLP) analysis (Clinithink with BioME)**

A Recent Article You May Enjoy

ANNUAL REVIEWS

Annual Review of Biomedical Data Science

Modern Clinical Text Mining: A Guide and Review

Bethany Percha

Department of Medicine and Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY 10025, USA; email: bethany.percha@mssm.edu

Annu. Rev. Biomed. Data Sci. 2021. 4:165–87

The *Annual Review of Biomedical Data Science* is online at biodatasci.annualreviews.org

https://doi.org/10.1146/annurev-biodatasci-030421-030931

Keywords

text mining, natural language processing, electronic health record, clinical text, machine learning

Abstract

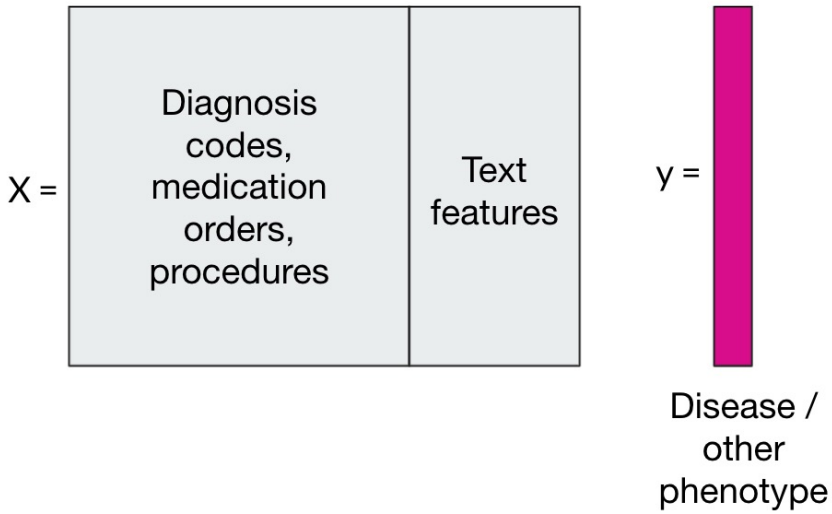
Electronic health records (EHRs) are becoming a vital source of data for healthcare quality improvement, research, and operations. However, much of the most valuable information contained in EHRs remains buried in unstructured text. The field of clinical text mining has advanced rapidly in recent years, transitioning from rule-based approaches to machine learning and, more recently, deep learning. With new methods come new challenges, however, especially for those new to the field. This review provides an overview of clinical text mining for those who are encountering it for the first time (e.g., physician researchers, operational analytics teams, machine learning scientists from other domains). While not a comprehensive survey, this review describes the state of the art, with a particular focus on new tasks and methods developed over the past few years. It also identifies key barriers between these remarkable technical advances and the practical realities of implementation in health systems and in industry.

Part II

Defining the Task and Approach

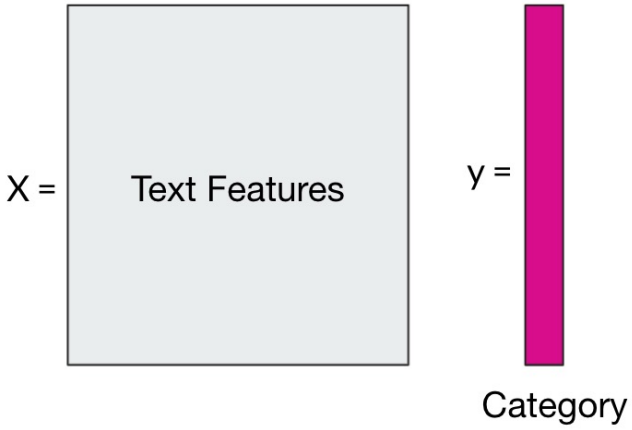
Information Extraction vs. Modeling

Clinical text can play multiple roles in a project, so it is important to start by defining one's overall goal and how the text fits in. Some projects combine text data with other features, while others are about interpreting the text itself.



Example: Electronic phenotyping

EHR search indexing, knowledgebase construction, patient timeline building likewise focus on IE



Examples: Classify mammography reports by BI-RADS category, classify pathology reports by diagnosis (also many unsupervised examples)

Rule-Based vs. Statistical

Rule-based systems codify expert knowledge into a set of structured rules, or templates, that produce structured information when applied to unstructured text. Statistical (ML) methods can be difficult in the clinical domain because of privacy concerns and lack of training data.

	Rule-based grammar	Machine Learning algorithm
+	★ Flexible ★ Easy to debug ★ Doesn't require a massive training corpus ★ Understanding of the language phenomenon ★ High precision	★ Easy to scale ★ "Learnability" without being explicitly programmed ★ Fast development (if datasets available) ★ High recall (coverage)
−	★ Requires skilled developers and linguists ★ Slow parser development ★ Moderate recall (coverage)	★ Requires training corpus with annotation ★ Difficult to debug ★ No understanding of the language phenomenon

https://medium.com/friendly-data/machine-learning-vs-rule-based-systems-in-nlp-5476de53c3b8

Part III
Clinical Information Extraction

Three Common Tasks

Because of the broad need for basic information extraction in applied tasks like medical coding, search, and case finding, software systems have been developed to perform these tasks automatically.

1. **Named Entity Recognition (NER)**
   Identifying and locating mentions of conceptual categories, such as drug, symptom, or disease names, in text

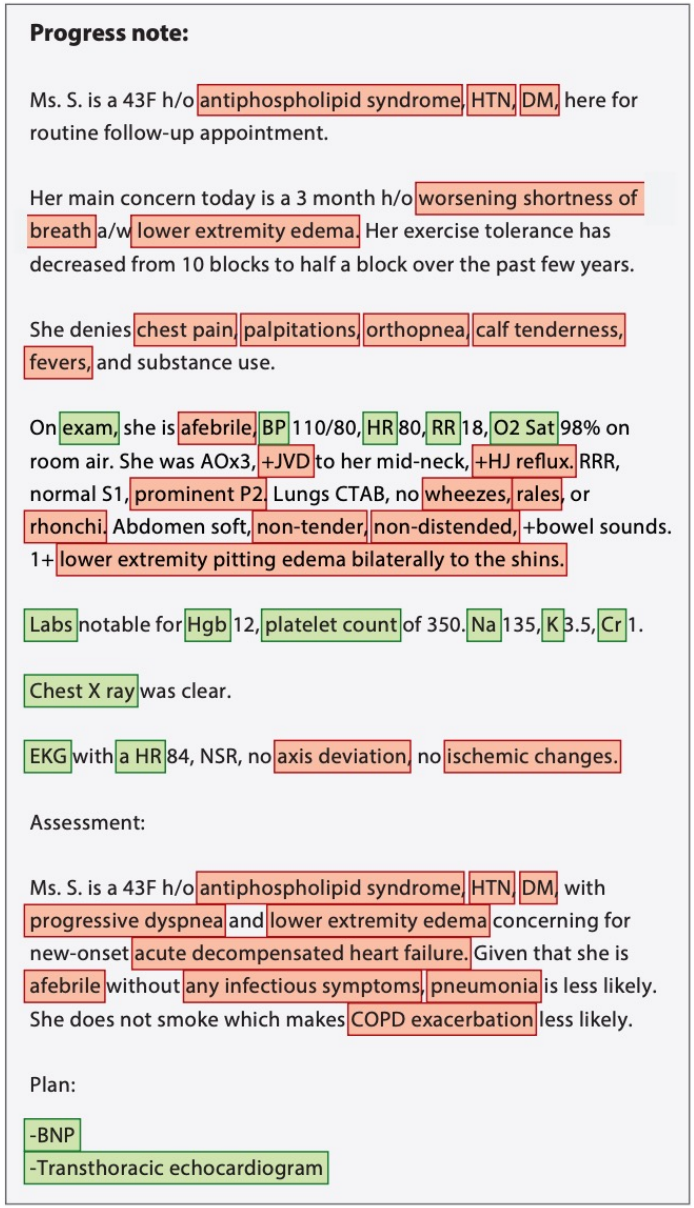
2. **Concept Normalization**
   Assigning a unique identity to an entity name recognized in text; in the biomedical domain, names are typically mapped to concepts from structured terminologies or ontologies

3. **Relation Extraction**
   Assigning a structured form to a relationship between entities; typically, this form includes the normalized entities and a label denoting the nature of their relationship

Named Entity Recognition (NER)

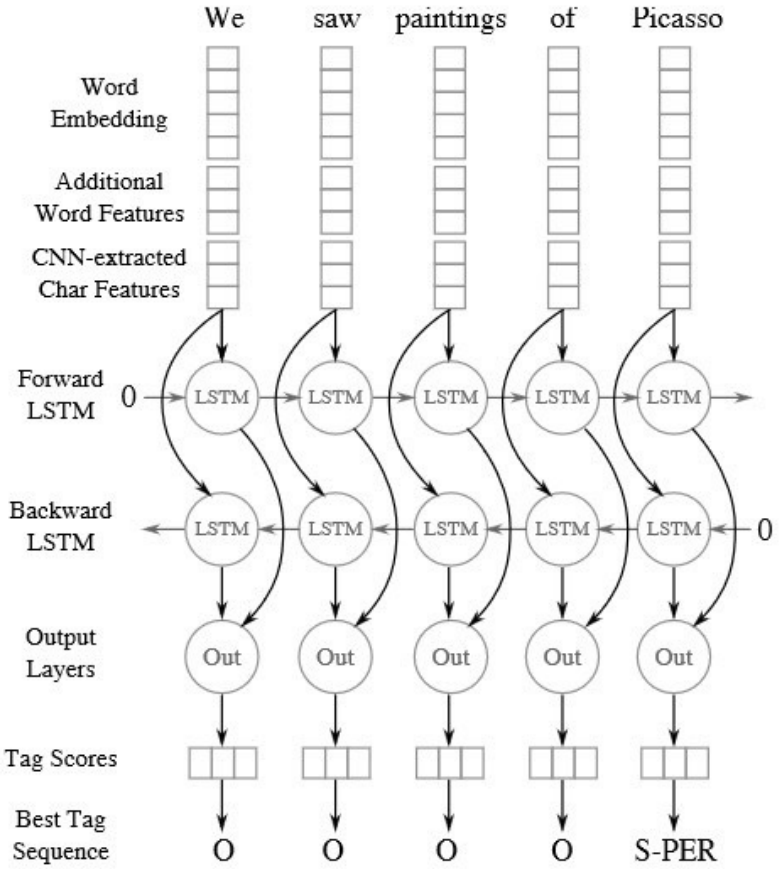
The annotations shown here are from the Stanza clinical text processing pipeline, trained using data from the 2010 i2b2/VA challenge.

What do you notice?



Methods for NER

Systems range from simple dictionary-based approaches to machine learning models adapted for sequence data, including conditional random fields (CRFs), recurrent neural networks (RNNs), and RNN variants such as long short-term memory networks (LSTMs).



https://towardsdatascience.com/named-entity-recognition-ner-meeting-industrys-requirement-by-applying-state-of-the-art-deep-698d2b3b4ede

Side Note…

Features for NER Algorithms

Tokenization

Tokenization is the act of splitting a text string into a list of tokens (e.g. words and punctuation, characters, etc.). The fastest tokenizers are based on regular expressions. For many applications, whitespace tokenization works just fine.

Original sentence:

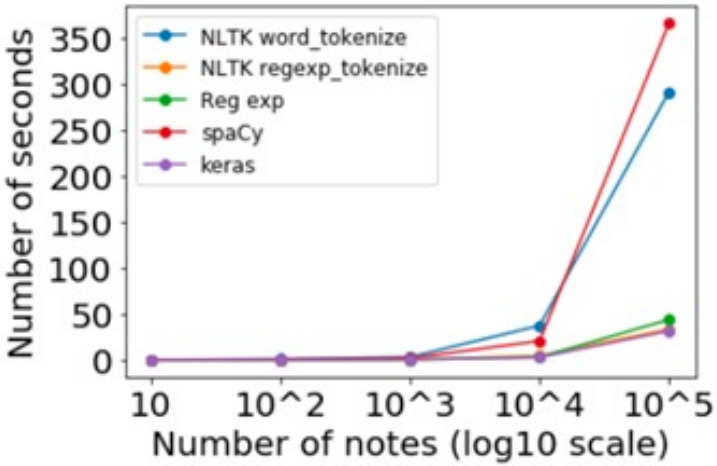
```
"Ethanol consumption reduces
endothelium-dependent relaxation
induced by acetylcholine."
```

Output of NLTK's word_tokenize:

```
"Ethanol", "consumption",
"reduces", "endothelium-dependent",
"relaxation", "induced", "by",
"acetylcholine", "."
```

Output of NLTK's wordpunct_tokenize:

```
"Ethanol", "consumption",
"reduces", "endothelium-dependent",
"relaxation", "induced", "by",
"acetylcholine", "."
```

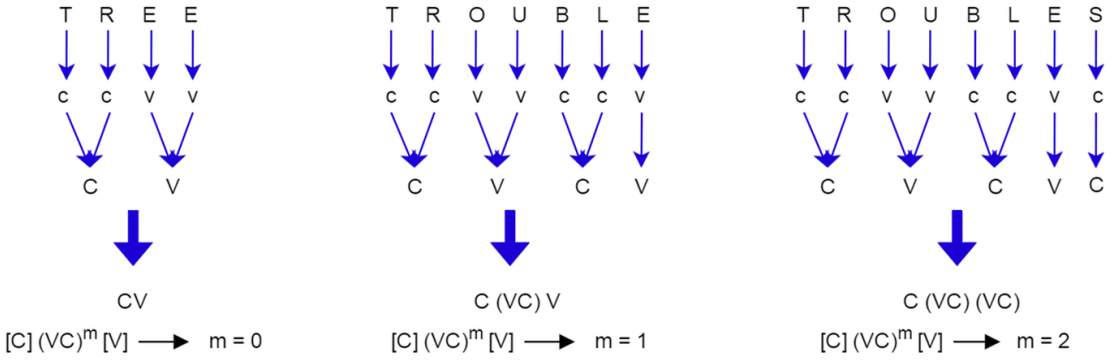


Source: https://towardsdatascience.com/benchmarking-python-nlp-tokenizers-3ac4735100c5

Stemming and Lemmatization

"The goal of both stemming and lemmatization is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form."

– Introduction to Information Retrieval (Manning and Schutze) –



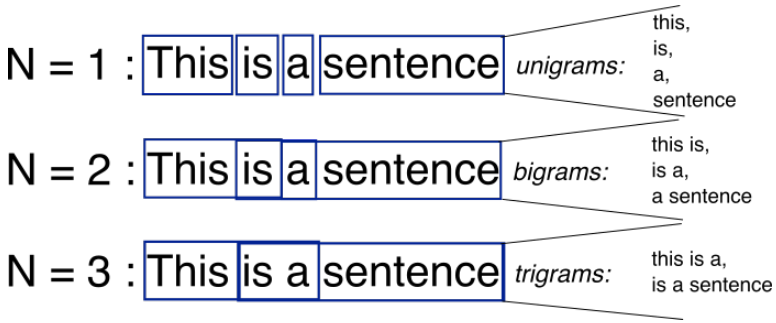
Porter Stemming Algorithm

SSES	→	SS	(m>0) ATIONAL	→	ATE
IES	→	I	(m>0) TIONAL	→	TION
SS	→	SS	(m>0) ENCI	→	ENCE
S	→		(m>0) ANCI	→	ANCE

Source: https://vijinimallawaarachchi.com/2017/05/09/porter-stemming-algorithm/

N-grams

A classic way of featurizing documents: collect all words and phrases of length N=1, 2, 3, etc. The combinatorics typically become a problem above N=3 (sometimes even N=2). N-grams can also be produced at the character level.



- unigram (1-gram):

a	swimmer	likes	swimming	thus	he	swims

- bigram (2-gram):

a swimmer	swimmer likes	likes swimming	swimming thus	...

- trigram (3-gram):

a swimmer likes	swimmer likes swimming	likes swimming thus	...

Sources: https://stackoverflow.com/questions/18193253/what-exactly-is-an-n-gram, https://sebastianraschka.com/images/blog/2014/naive_bayes_1/

TFIDF

How should one select the top features for downstream use in classification algorithms? It's not good to use pure frequency because the most frequent N-grams are often the least informative.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

Sentence 1: The car is driven on the road.
Sentence 2: The truck is driven on the highway.

Word	TF		IDF	TF*IDF	
	A	B		A	B
The	1/7	1/7	log(2/2) = 0	0	0
Car	1/7	0	log(2/1) = 0.3	0.043	0
Truck	0	1/7	log(2/1) = 0.3	0	0.043
Is	1/7	1/7	log(2/2) = 0	0	0
Driven	1/7	1/7	log(2/2) = 0	0	0
On	1/7	1/7	log(2/2) = 0	0	0
The	1/7	1/7	log(2/2) = 0	0	0
Road	1/7	0	log(2/1) = 0.3	0.043	0
Highway	0	1/7	log(2/1) = 0.3	0	0.043

Source: https://www.freecodecamp.org/news/how-to-process-textual-data-using-tf-idf-in-python-cd2bbc0a94a3/

Feature Selection and Normalization

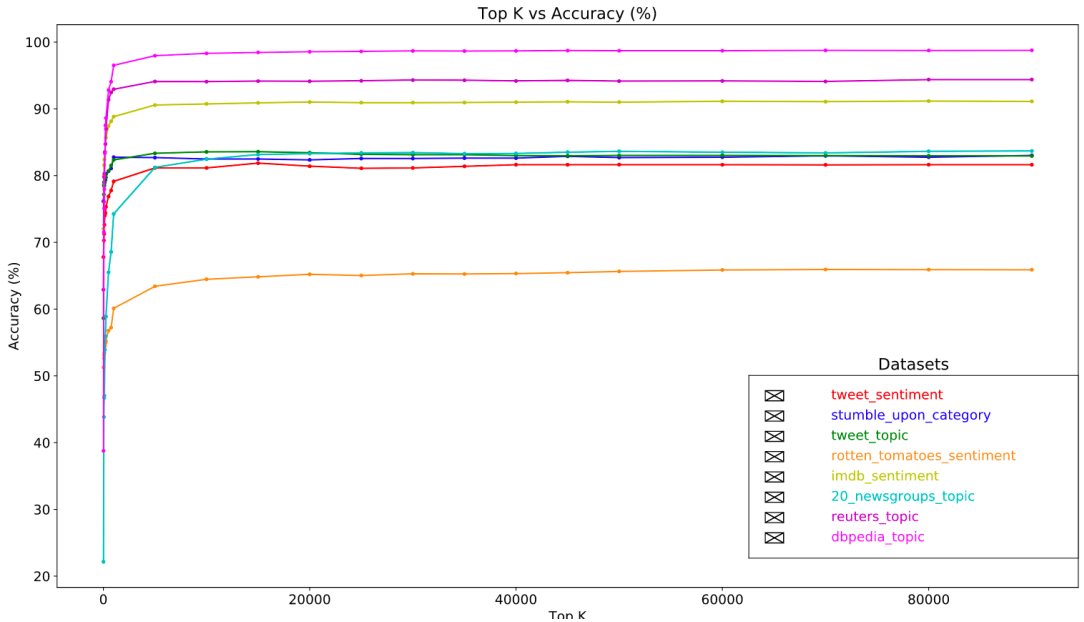
Text classification accuracy tends not to benefit much from features beyond the top ~20k. You may choose to **normalize** features by sample or by feature or do nothing (most common).

`sklearn.feature_selection`: Feature Selection

The `sklearn.feature_selection` module implements feature selection algorithms. It currently includes univariate filter selection methods and the recursive feature elimination algorithm.

User guide: See the Feature selection section for further details.

`feature_selection.GenericUnivariateSelect`([...])	Univariate feature selector with configurable strategy.
`feature_selection.SelectPercentile`([...])	Select features according to a percentile of the highest scores
`feature_selection.SelectKBest`([score_func, k])	Select features a
`feature_selection.SelectFpr`([score_func, alpha])	Filter: Select the
`feature_selection.SelectFdr`([score_func, alpha])	Filter: Select the
`feature_selection.SelectFromModel`(estimator)	Meta-transforme weights.
`feature_selection.SelectFwe`([score_func, alpha])	Filter: Select the
`feature_selection.RFE`(estimator[, ...])	Feature ranking v
`feature_selection.RFECV`(estimator[, step, ...])	Feature ranking v validated selectic
`feature_selection.VarianceThreshold`([threshold])	Feature selector
`feature_selection.chi2`(X, y)	Compute chi-squa and class.
`feature_selection.f_classif`(X, y)	Compute the ANC
`feature_selection.f_regression`(X, y[, center])	Univariate linear re
`feature_selection.mutual_info_classif`(X, y)	Estimate mutual in
`feature_selection.mutual_info_regression`(X, y)	Estimate mutual in

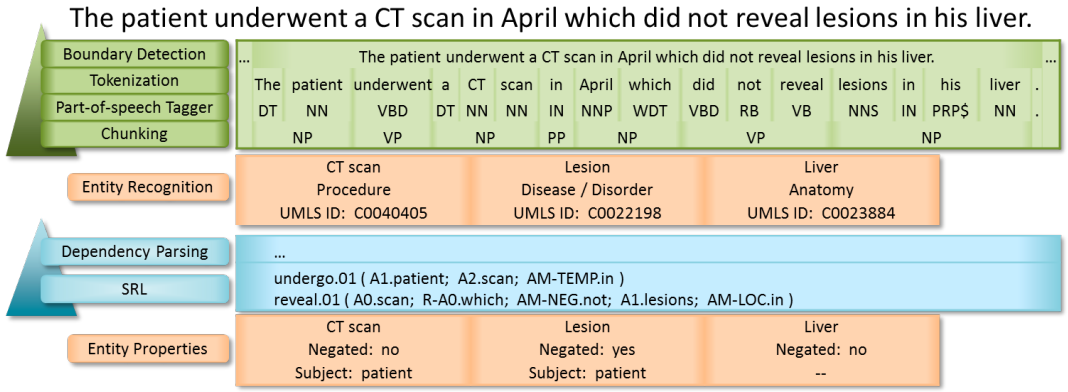


Source: https://developers.google.com/machine-learning/guides/text-classification/step-3

Biomedical Feature Extraction and UMLS

There are a variety of biomedical text annotators available that will extract biomedically or clinically relevant concepts from free text. Most of these map text strings to concepts from the Unified Medical Language System (UMLS). **They do not always improve classification performance.**

cTAKES
https://ctakes.apache.org/



MetaMapLite
https://metamap.nlm.nih.gov/MetaMapLite.shtml

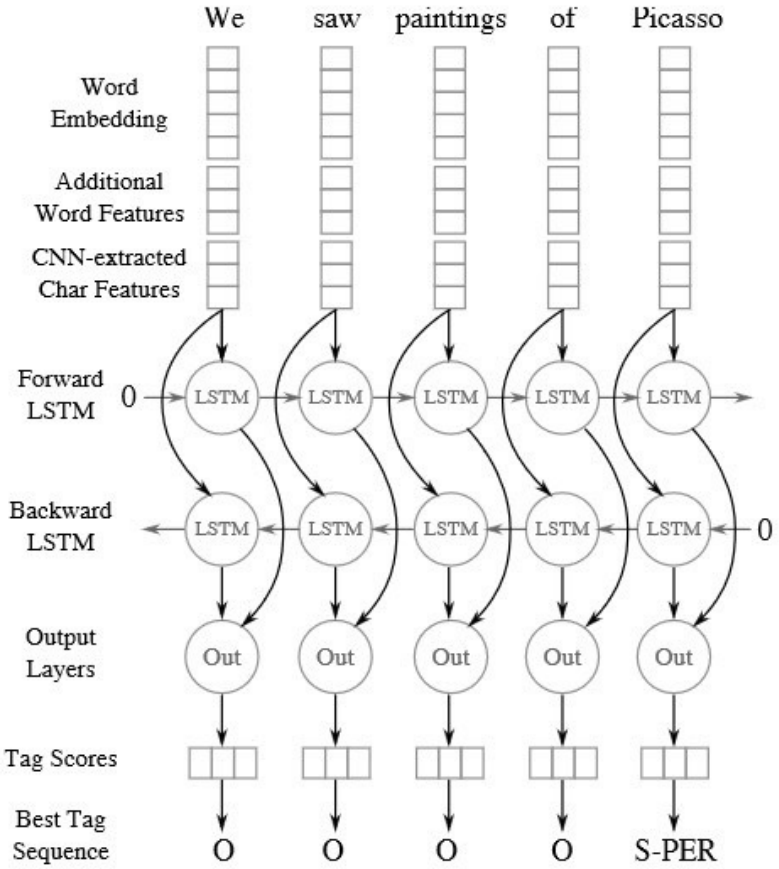
Sentences	Detected phrases	Extracted concepts	Selected
First sentence	Hyperlipidaemia	**[Disease or Syndrome]**	✓
		[Finding]	×
	Patient	[Patient or Disabled group]	×
	Lipitor	[Organic Chemical, Pharmacologic Substance]	×
	80%	[Quantitative Concept]	×
	mg++ increased	[Finding]	×
Second sentence	Progress note	[Clinical Attribute]	×
		[Intellectual Product]	×
	Patient chart	[Manufactured Object]	×
	Assisted living facility	[Healthcare Related Organization, Manufactured Object]	×
	Patient	[Patient or Disabled group]	×
	Shortness of breath	**[Sign or Symptom]**	✓

Source: Abdollahi et al. (2018). Uncovering Discriminative Knowledge-Guided Medical Concepts for Classifying Coronary Artery Disease Notes: 31st Australasian Joint Conference, Wellington, New Zealand, December 11-14, 2018.

Back to Part III…

Methods for NER (review)

Systems range from simple dictionary-based approaches to machine learning models adapted for sequence data, including conditional random fields (CRFs), recurrent neural networks (RNNs), and RNN variants such as long short-term memory networks (LSTMs).



https://towardsdatascience.com/named-entity-recognition-ner-meeting-industrys-requirement-by-applying-state-of-the-art-deep-698d2b3b4ede

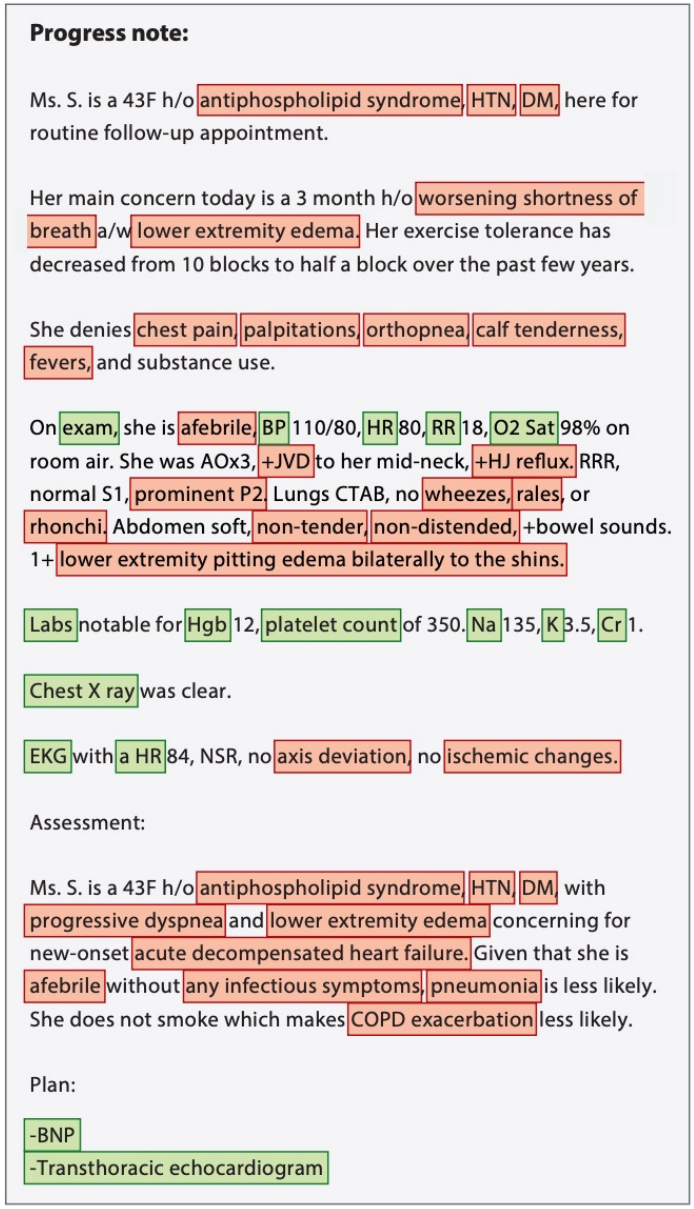
Domain Specificity and Key Challenges

NER only makes sense when the entities involved are discrete and have defined locations in text.

Not all NER annotations are correct or meaningful without consideration of the surrounding context.

One may be interested in an entity class for which no pre-annotated corpus or pretrained model is available.

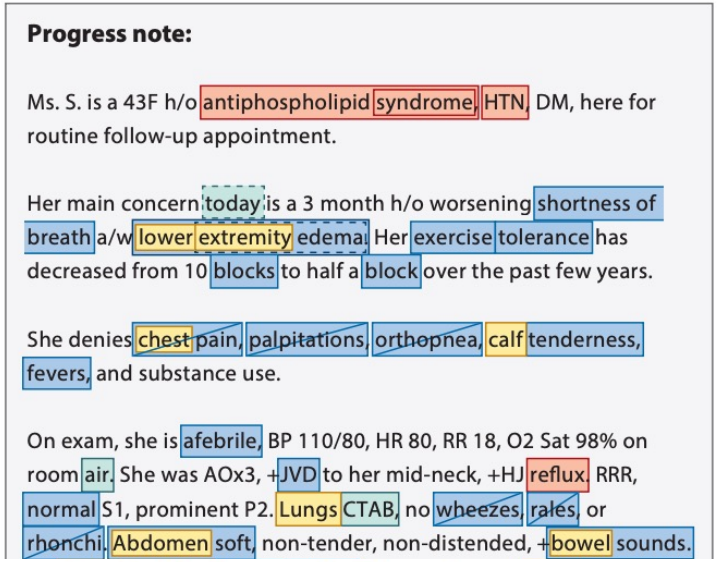
General domain NER models trained using general domain corpora – may not work on clinical text.



Concept Normalization

Also known as “entity linking”, this is the task of assigning a unique identity to each entity name mentioned in a text.

What do you notice?



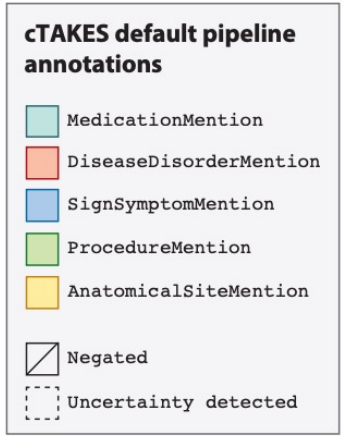


Table 2 Examples of cTAKES annotations associated with the note in Figure 2

Line number(s)	Annotation type	Original string	Normalized term	UMLS concept ID
2	DiseaseDisorderMention	HTN	Hypertensive disease	C0020538
4–5	SignSymptomMention	shortness of breath	Dyspnea	C0013404
7	SignSymptomMention	chest pain (negated)	Chest pain (negated)	C0008031
10	SignSymptomMention	JVD	Jugular venous engorgement	C0425687
16	ProcedureMention	EKG	Electrocardiography	C1623258
24	ProcedureMention	Transthoracic echocardiogram	Transthoracic echocardiography	C0430462
10	DiseaseDisorderMention	reflux	Gastresophageal reflux disease	C0017168
11	MedicationMention	CTAB	Cetrimonium bromide	C0951233
23	MedicationMention	BNP	Nesiritide	C0054015

Numbers, Ranges, and Sections

Identifying and normalizing numeric values and identifying note sections are tasks somewhat unique to clinical text. To date, the SecTag system developed by Denny et al is the only section identification system used outside the institution in which it was developed.

Research article | Open Access | Published: 18 July 2019

Current approaches to identify sections within clinical narratives from electronic health records: a systematic review

Alexandra Pomares-Quimbaya ✉, Markus Kreuzthaler & Stefan Schulz

BMC Medical Research Methodology **19**, Article number: 155 (2019) | Cite this article

3864 Accesses | **2** Citations | **1** Altmetric | Metrics

Abstract

Background

The identification of sections in narrative content of Electronic Health Records (EHR) has demonstrated to improve the performance of clinical extraction tasks; however, there is not yet a shared understanding of the concept and its existing methods. The objective is to report the results of a systematic review concerning approaches aimed at identifying sections in narrative content of EHR, using both automatic or semi-automatic methods.

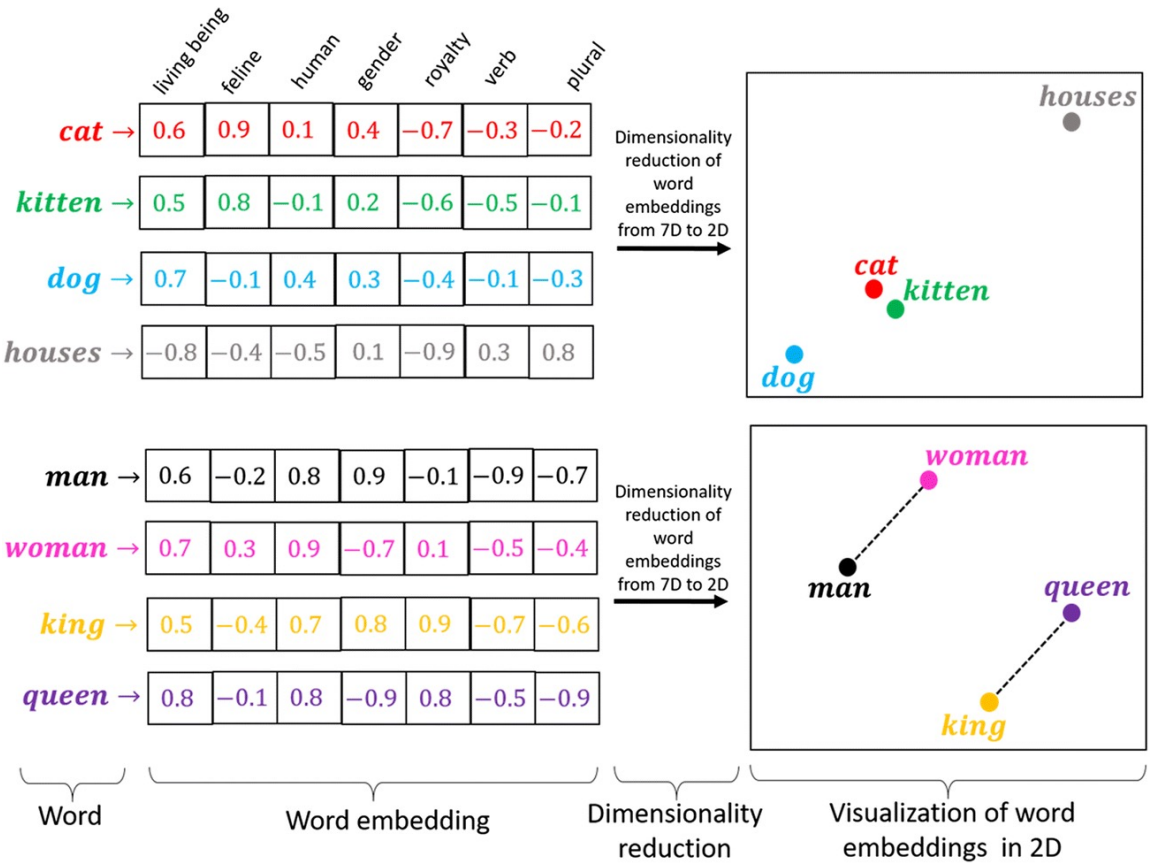
Methods

This review includes articles from the databases: SCOPUS, Web of Science and PubMed (from January 1994 to September 2018). The selection of studies was done using predefined

Part IV
Embeddings and Pretraining

Embeddings

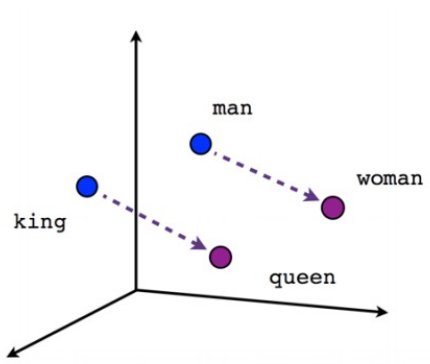
An embedding is a semantically meaningful mathematical representation of a word, phrase, or other piece of text. Usually a vector, it is designed in such a way that words and phrases with similar meanings have similar vectors.



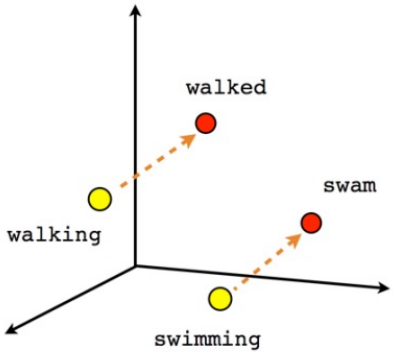
https://swatimeena989.medium.com/training-word2vec-using-gensim-14433890e8e4

Embeddings are not a New Idea

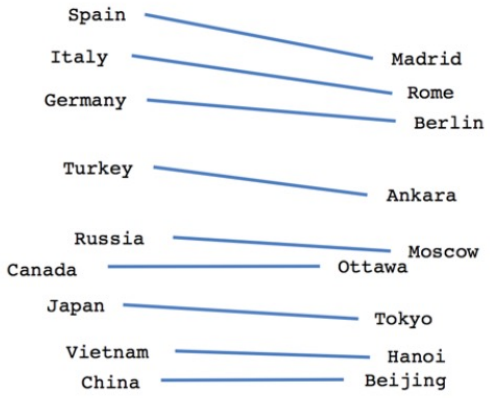
The field of distributional semantics originated with Latent Semantic Indexing in the late 1980s and reached a milestone with the development of word2vec and GloVe in 2013-2014. It is a collection of methods that create vector space embeddings of words and phrases that reflect how they are used in context.



Male-Female



Verb tense



Country-Capital

Why are embeddings useful?

(1) Do not require annotated corpora for training, so it is easy to create embeddings that are specific to clinical text or that capture regularities of expression within a clinical subfield or institution. (2) An embedding can incorporate structured information beyond what is found in text, and embeddings have been created to represent CUIs, documents, or entire patient records.

Journal of Biomedical Informatics 101 (2020) 103323

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

SECNLP: A survey of embeddings in clinical natural language processing

Katikapalli Subramanyam Kalyan*, S. Sangeetha

Text Analytics and NLP Lab, Department of Computer Applications, NIT Trichy, India

ARTICLE INFO

Keywords:
Embeddings
Distributed representations
Medical
Natural language processing
Survey

ABSTRACT

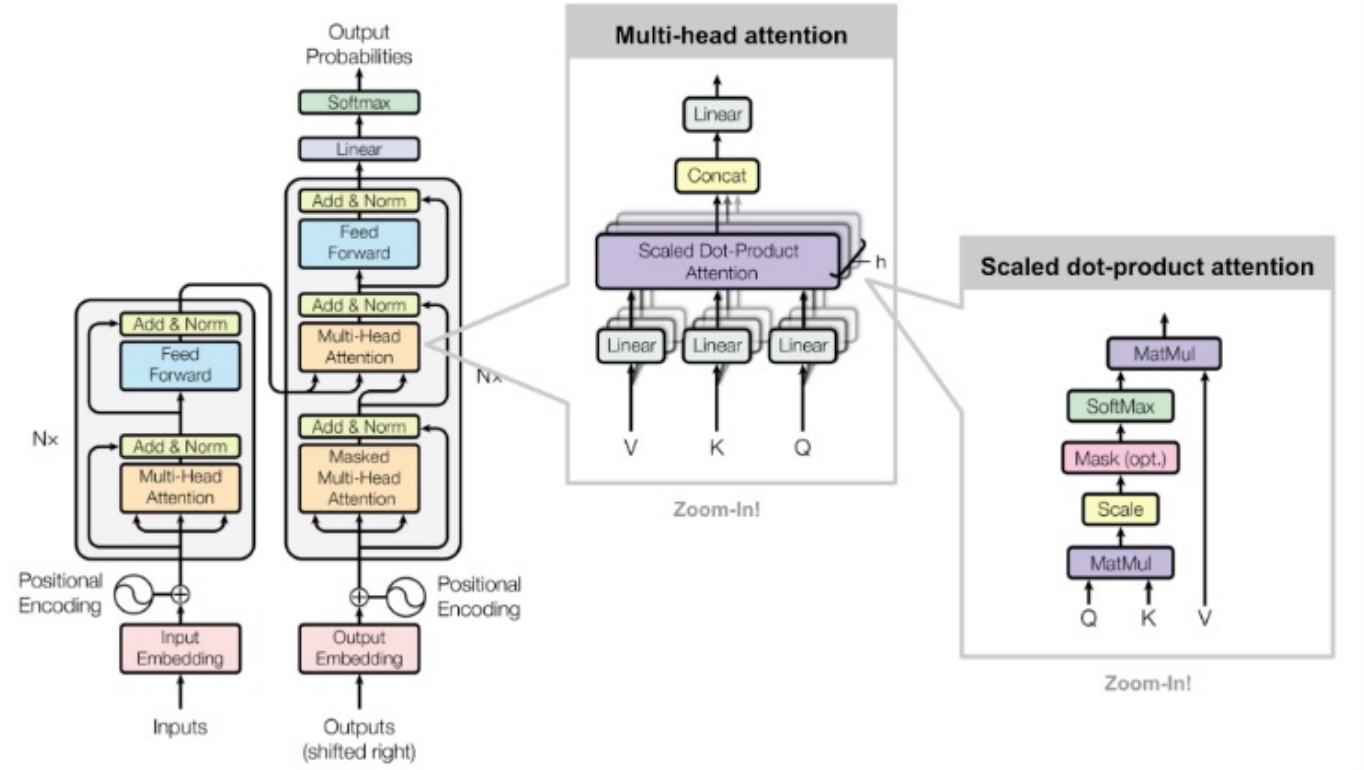
Distributed vector representations or embeddings map variable length text to dense fixed length vectors as well as capture prior knowledge which can transferred to downstream tasks. Even though embeddings have become de facto standard for text representation in deep learning based NLP tasks in both general and clinical domains, there is no survey paper which presents a detailed review of embeddings in Clinical Natural Language Processing. In this survey paper, we discuss various medical corpora and their characteristics, medical codes and present a brief overview as well as comparison of popular embeddings models. We classify clinical embeddings and discuss each embedding type in detail. We discuss various evaluation methods followed by possible solutions to various challenges in clinical embeddings. Finally, we conclude with some of the future directions which will advance research in clinical embeddings.

1. Introduction

1.1. Literature Selection

Contextual Embeddings and Pretraining

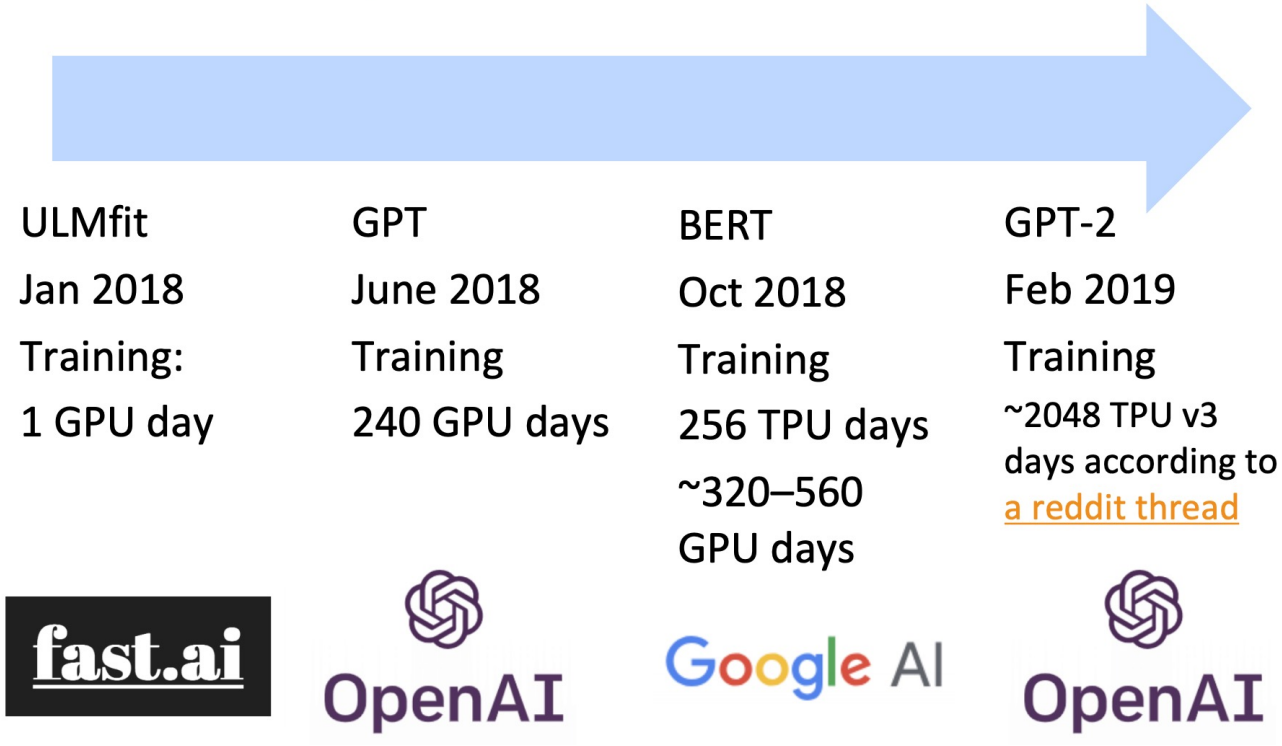
Novel neural network architectures have permitted the creation of embeddings that vary depending on the context. This has expanded the representational power of embedding methods.



https://medium.com/@shreyasikalra25/predict-movie-reviews-with-bert-88d8b79f5718

Transformer Architectures

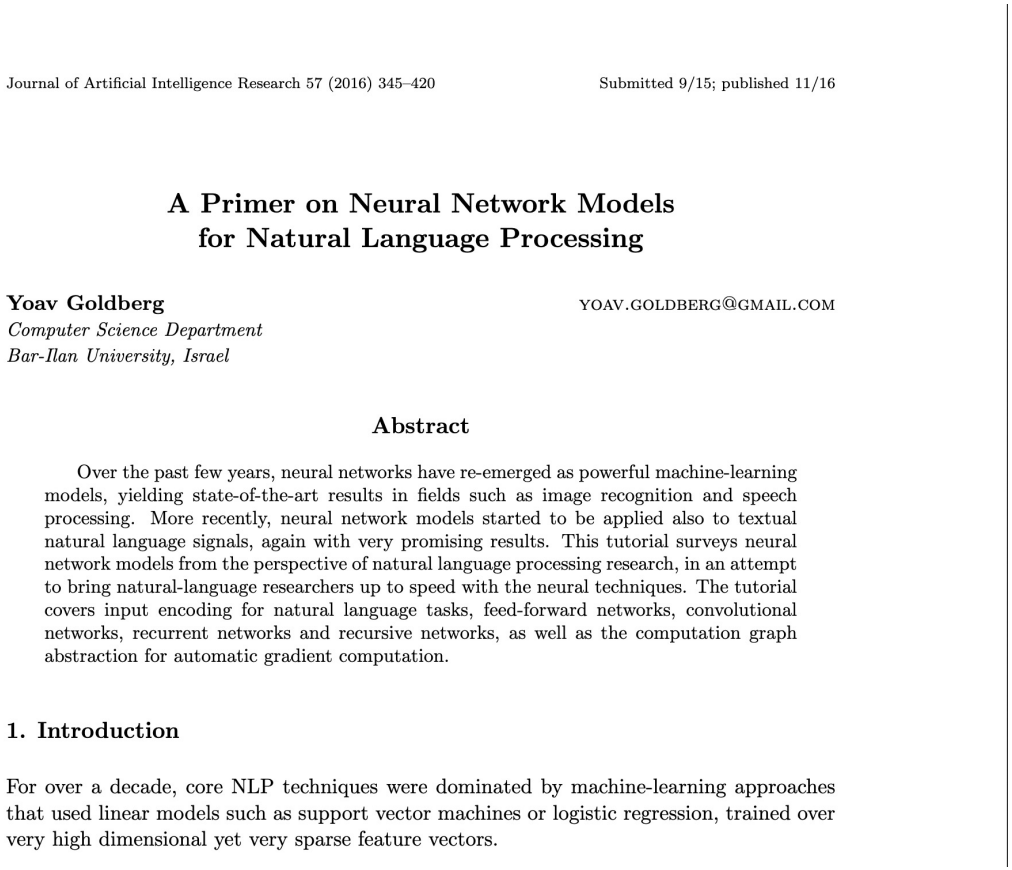
The best-known Transformer-based model is called BERT. It was developed by a team at Google and published in 2019.



Source: Course notes for CS224N (Stanford). Lecture 14: Contextual Representations and Pretraining

A Good Place to Start

If you really want to understand neural network methods for natural language processing, read this paper first (and then study all the transformer-based models that have come out in the last three years).



Journal of Artificial Intelligence Research 57 (2016) 345–420 Submitted 9/15; published 11/16

A Primer on Neural Network Models for Natural Language Processing

Yoav Goldberg YOAV.GOLDBERG@GMAIL.COM
Computer Science Department
Bar-Ilan University, Israel

Abstract

Over the past few years, neural networks have re-emerged as powerful machine-learning models, yielding state-of-the-art results in fields such as image recognition and speech processing. More recently, neural network models started to be applied also to textual natural language signals, again with very promising results. This tutorial surveys neural network models from the perspective of natural language processing research, in an attempt to bring natural-language researchers up to speed with the neural techniques. The tutorial covers input encoding for natural language tasks, feed-forward networks, convolutional networks, recurrent networks and recursive networks, as well as the computation graph abstraction for automatic gradient computation.

1. Introduction

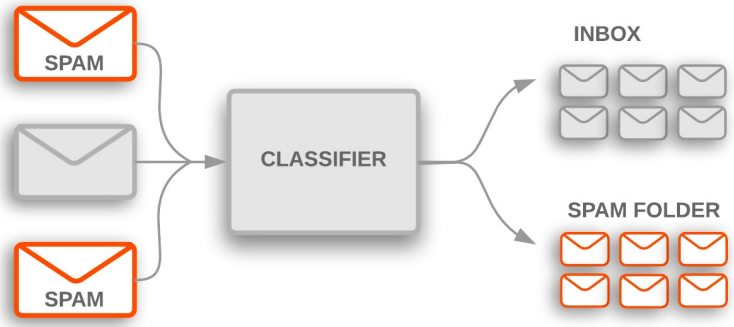
For over a decade, core NLP techniques were dominated by machine-learning approaches that used linear models such as support vector machines or logistic regression, trained over very high dimensional yet very sparse feature vectors.

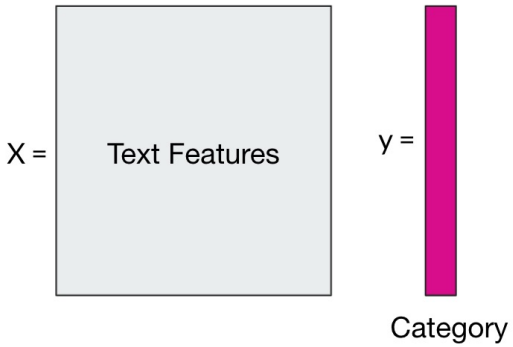
http://jalammar.github.io/illustrated-bert/

Part V
Text Classification

Text Classification

The goal is to classify the text of a document (where a document can be as short as a single phrase or sentence and as long as a book) into two or more categories. A recent survey found that of 212 recent clinical text mining papers, 88 (41.5%) focused on text classification.





Examples:

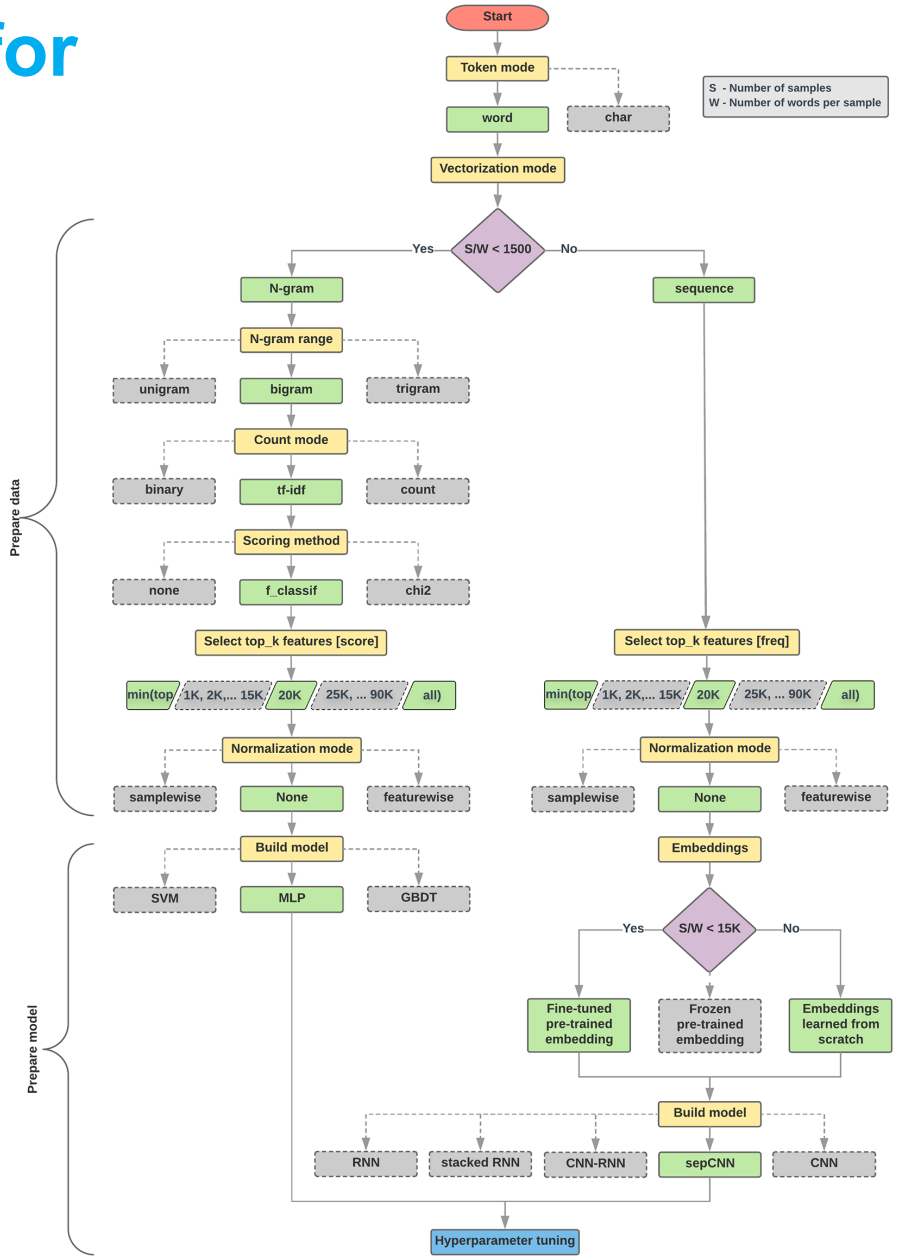
- Classifying primary care descriptions of back pain into acute vs. chronic
- Distinguishing normal vs. abnormal knee MRI reports
- Assessing whether a patient is a current or former smoker based on clinical notes

General Procedure for Text Classification

Any machine learning classifier can be adapted to classify text.

The hard part is figuring out how to turn a document into a set of features suitable for classification.

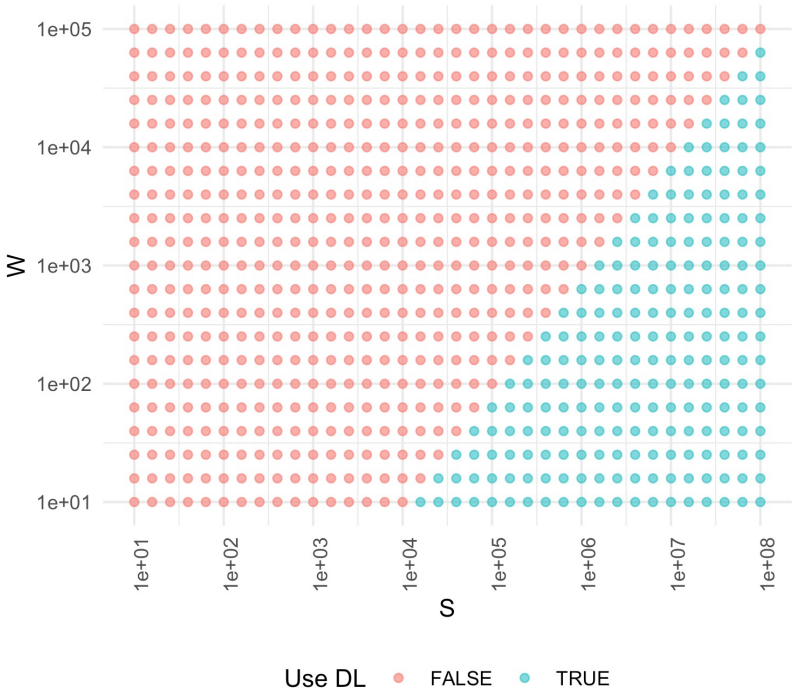
Google provides this handy guide

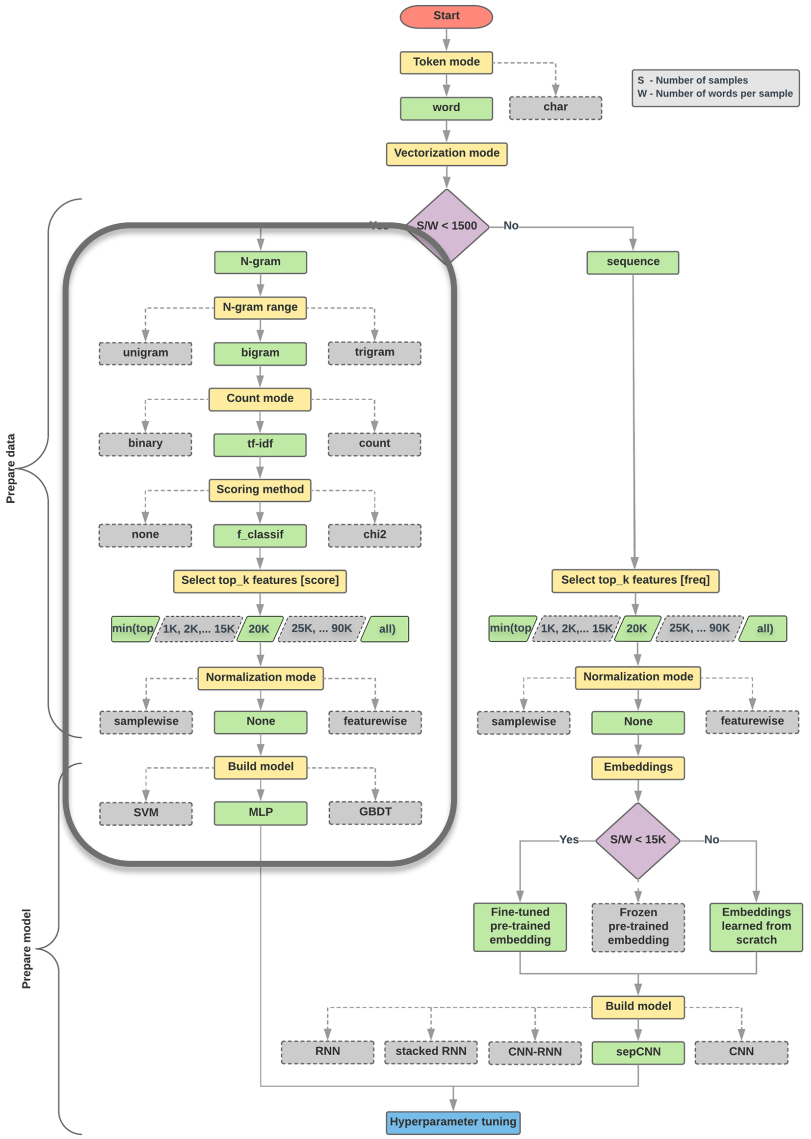


https://developers.google.com/machine-learning/guides/text-classification

The Standard Pipeline

Start with raw text, tokenize it, extract N-grams, weight them somehow (or not), select top features, normalize them (or not), use whatever supervised ML algorithm you want to classify them.

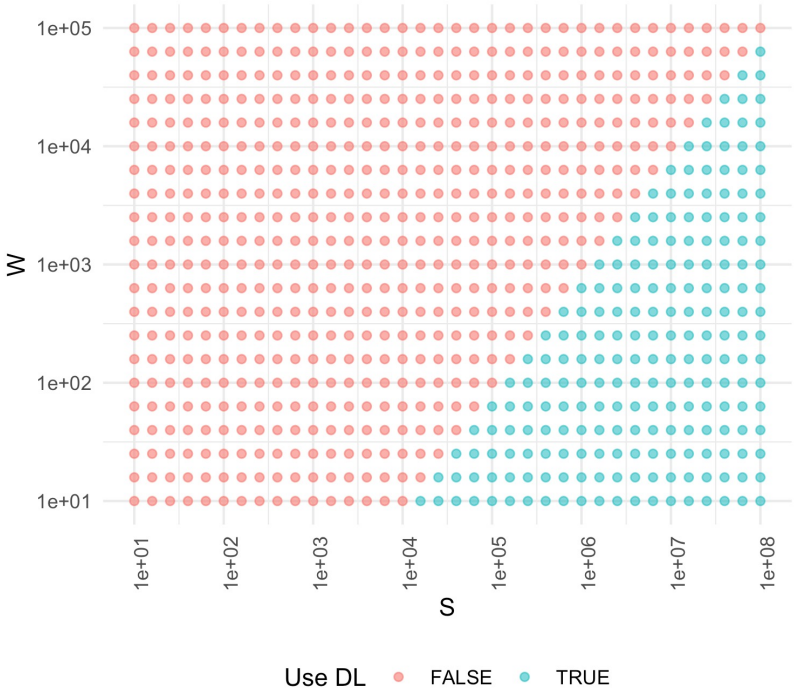


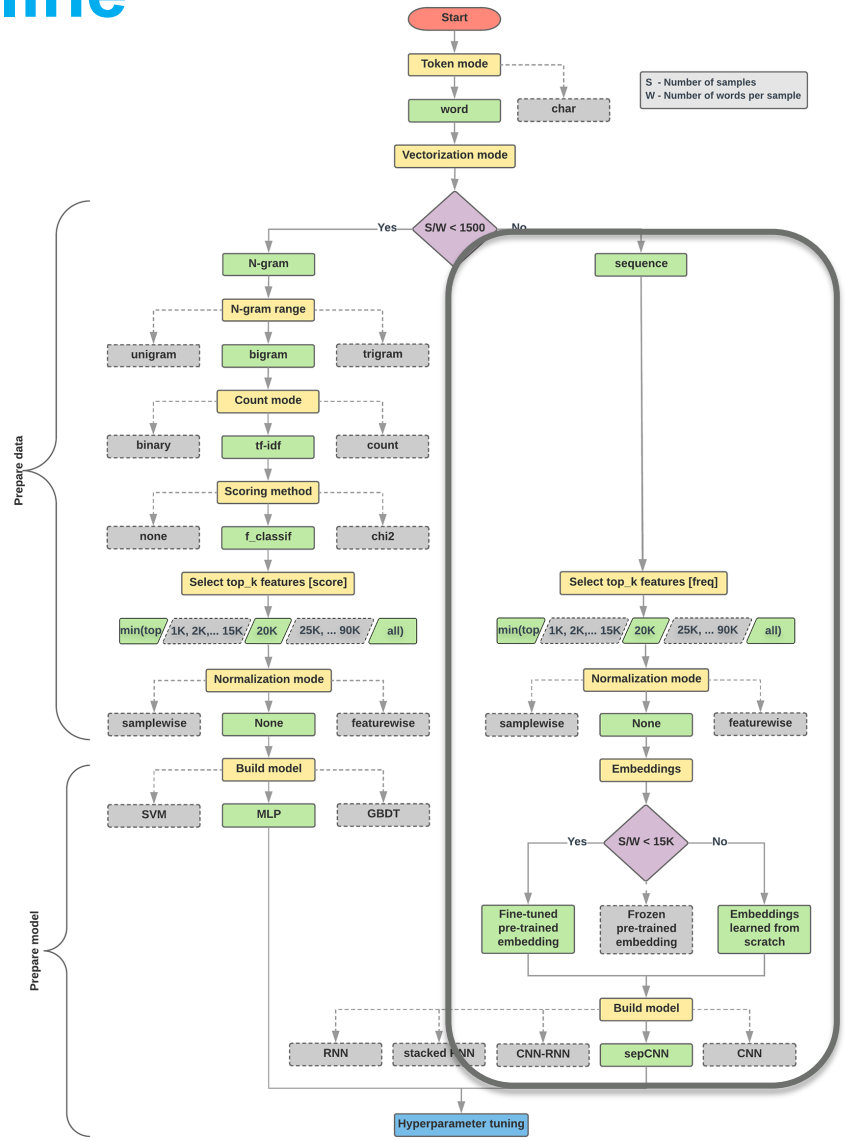


The Deep Learning Pipeline

Start with raw text, tokenize it, do feature selection (maybe), produce embeddings or use pretrained embeddings, feed to neural network.

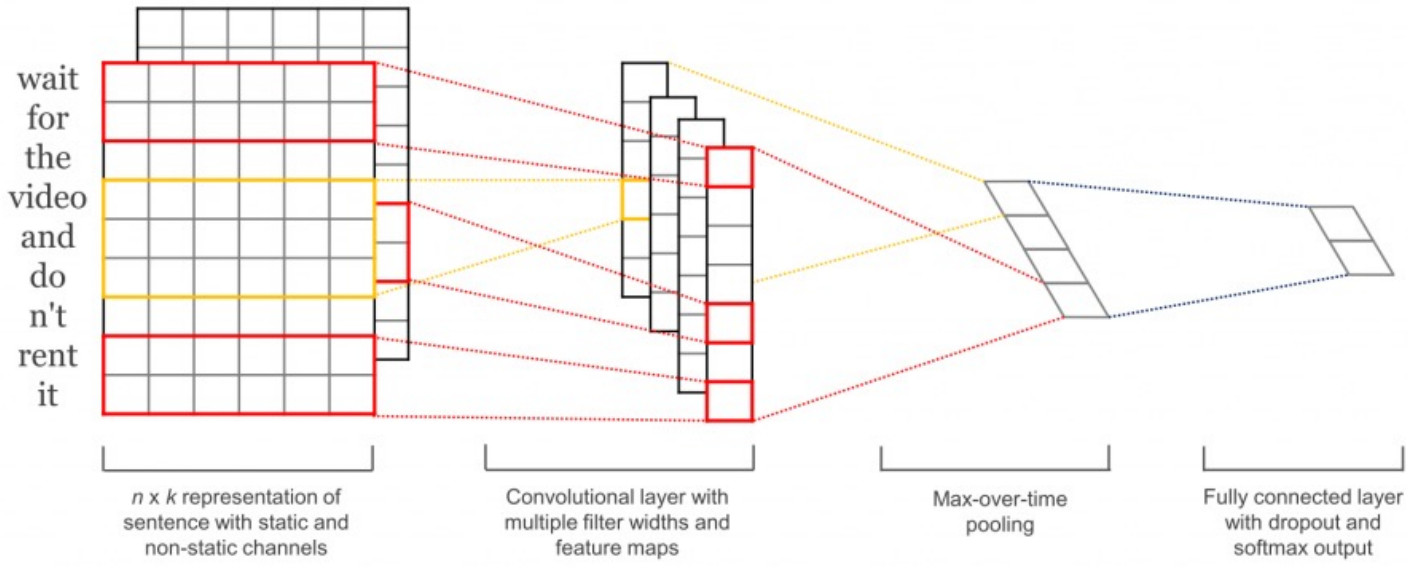
Convolutional neural networks (CNNs) are most popular architecture at the moment.





Text Classification with CNNs

These models have been deployed on a variety of tasks, including assigning diagnosis codes, classifying radiology reports, subtyping diseases, and determining the presence or absence of comorbidities.



http://www.wildml.com/2015/12/implementing-a-cnn-for-text-classification-in-tensorflow/

Part VI

Additional Topics: Weak Supervision and Inference/Entailment

Weak and Distant Supervision

Machine learning approaches to clinical NLP generally suffer from a lack of training data. Existing clinical information extraction and text classification models have generally been trained using the same few annotated datasets, which restricts the range and quality of annotations they produce.

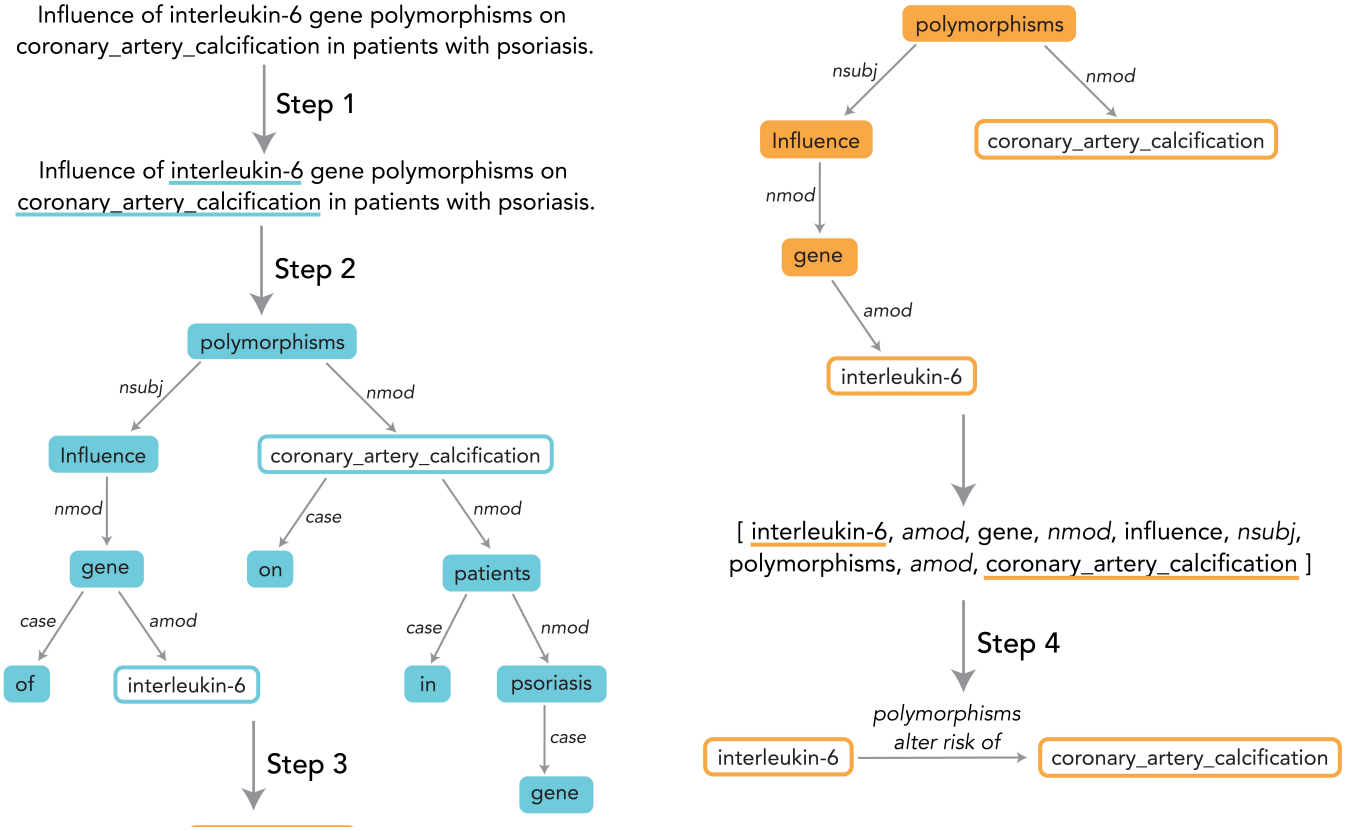
1. **Weak Supervision**
   Supervised learning using weak (noisy) labels; for example, simple heuristic rules (labeling functions) may be used to create large, weakly annotated training sets

2. **Distant Supervision**
   Supervised learning using training signals that do not directly label the training examples (e.g., using structured clinical data to train a text mining algorithm, since data associated with patients/encounters, not sentences/documents)

3. **Alternatives: Crowdsourcing and Active Learning**
   Crowdsourcing not usually a viable option due to privacy concerns. Active learning is a strategy for minimizing annotation effort by iteratively sampling subsets of data for human annotation based on current predictions of a supervised learning algorithm.

Relation Extraction

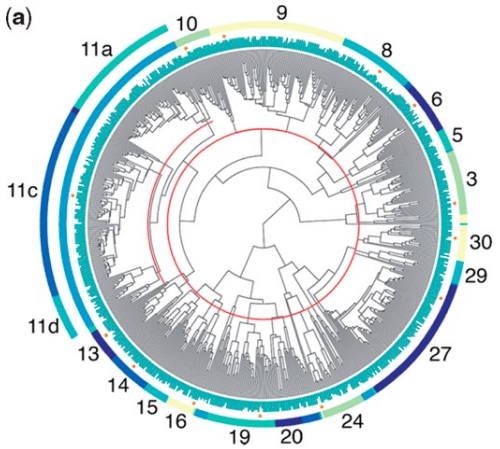
Relation extraction is the task of assigning a structured form to a relationship between or among entities based on how they are described in text. Typically, this form includes the categories of the entities and a label denoting the nature of their relationship.



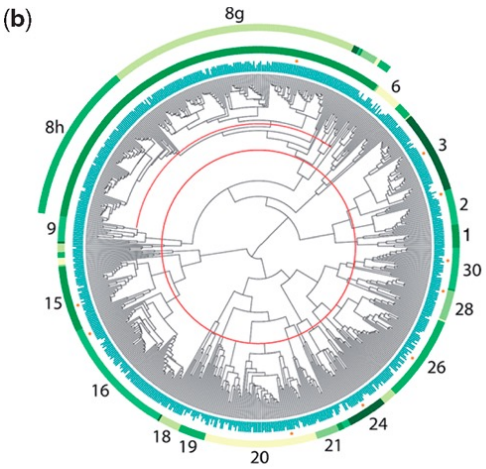
Percha B, Altman RB (2018) A global network of biomedical relationships derived from text. *Bioinformatics* 34(15), 2614-2624. Figure 1.

Example from the Biomedical Domain: GNBR

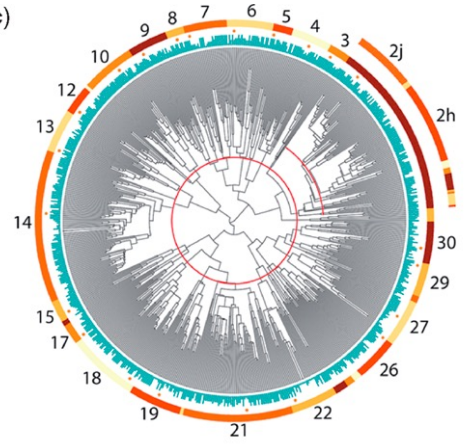
Cluster patterns from text (dependency paths) into semantically meaningful categories.



ID	Size	Description	Dependency Path Examples (represented as phrases)
3	36	inhibition	"C, a G inhibitor" "G specific inhibitor, C" "C, an inhibitor of G" "G inhibition by C" "effects of the G inhibitor, C, on…"
11c	96	metabolism, secretion/ uptake	"effect of G on C metabolism" "effects of G on C formation" "effect of G on the secretion of C" "control of G by C" "G stimulates C uptake" "[chemical] may reduce G concentration via C" "G stimulates C transport" "effect of C on G release"
19	38	channels / transporters	"regulation of G transporters C and…" "G is a C channel that modulates…" "G C transporter expression" "C transporter, G" "distribution of G C channel subunits"



ID	Size	Description	Dependency Path Examples (represented as phrases)
8h	80	treatment of disease (indication of efficacy)	"C may be useful for the treatment of D" "evaluate the protective efficacy of C in D" "C is a promising treatment option for patients with D" "C is approved for the treatment of D" "C is commonly prescribed for D"
15	37	side effects (association)	"D associated with C therapy" "the use of C has been associated with D" "C intake was associated with D" "incidence of D in patients receiving C" "D occurred after C"
20	63	levels associated with disease risk / progression	"high C levels are associated with increased risk of D" "C implicated in D" "effect of D on serum C levels" "patients with D and increased C concentrations" "C has been implicated in the pathogenesis of D" "C intake may be associated with [lower/higher] risk of D" "C supplementation and incidence of D: …"



ID	Size	Description	Dependency Path Examples (represented as phrases)
2j	33	influences disease treatment	"the use of G in the treatment of D" "D in patients treated with G" "effect of G on [event] in D patients" "G therapy in patients with D" "efficacy of G in D"
7	26	biomarkers, diagnostic	"G is a robust diagnostic biomarker for D" "G is an independent predictor of D" "G as an indicator of D in patients with…" "prognostic significance of G in D patients" "effects of [situation/event] on G levels in D" "G is a potential marker of D"
14	91	causal mutations	"mutation of G in a patient with D" "G mutation is associated with D" "novel mutation in G gene associated with D" "characterization of G mutations causing D" "mutations of the G gene in patients with D" "D: a novel G mutation…" "D: novel G mutations and…" "the recurrent mutation of G in C patients" "G mutations can cause D"

Percha B, Altman RB (2018) A global network of biomedical relationships derived from text. *Bioinformatics* 34(15), 2614-2624. Figure 3.

Why Relation Extraction is Important

Addresses issue of compositionality, the combining of individual facts to generate composite ideas. Compositionality presents a particularly important challenge for clinical text mining because clinical writing reflects a high level of assumed knowledge, as well as unstated implications about the temporal and causal ordering of events.

Some Opinionated Commentary

- State-of-the-art machine learning models, which are very data hungry, are likely to hit a dead end when it comes to relation extraction and inference.
- Transfer learning with massive language models is a promising approach, but it is still basically pattern matching.
- Text mining is the key to unlocking the potential of electronic health records, esp. for generating “real world evidence”.
- There is a lot of money to be made if someone can crack this. Health technology companies often imply they are using NLP for data curation, but this is mostly a lie.

Great Reviews of the Field

A REVIEW OF REVIEWS

The field of clinical text mining has been extensively reviewed in prior articles. The reviews selected below are those I found to be particularly useful surveys of specific research areas or the field in general.

Year	Author(s)	Title	Reference
2011	Chapman *et al*	Overcoming Barriers to NLP for Clinical Text: The Role of Shared Tasks and the Need for Additional Creative Solutions	(141)
2016	Ford *et al*	Extracting Information from the Text of Electronic Medical Records to Improve Case Detection: A Systematic Review	(142)
2016	Koleck *et al*	Natural Language Processing of Symptoms Documented in Free-Text Narratives of Electronic Health Records: A Systematic Review	(5)
2017	Kreimeyer *et al*	Natural Language Processing Systems for Capturing and Standardizing Unstructured Clinical Information: A Systematic Review	(8)
2019	Khattak *et al*	A Survey of Word Embeddings for Clinical Text	(143)
2019	Mujtaba *et al*	Clinical Text Classification Research Trends: Systematic Literature Review and Open Issues	(18)
2020	Spasic *et al*	Clinical Text Data in Machine Learning: Systematic Review	(24)
2010	Stanfill *et al*	A Systematic Literature Review of Automated Clinical Coding and Classification Systems	(94)
2018	Velupillai *et al*	Using Clinical Natural Language Processing for Health Outcomes Research: Overview and Actionable Suggestions for Future Advances	(144)
2018	Wang *et al*	Clinical Information Extraction Applications: A Literature Review	(21)
2020	Wu *et al*	Deep Learning in Clinical Natural Language Processing: A Methodical Review	(23)

Questions?

bethany.percha@mssm.edu