Lessons from natural language inference in the clinical domain

Vincent Frappier

03/10/18

Mila Medical

Electronic Health Records

Clinical note

Homme, 24 ans, électricien, douleur au coude droit, consomme drogue régulièrement

prescription: Aspirin 100 mg au besoin

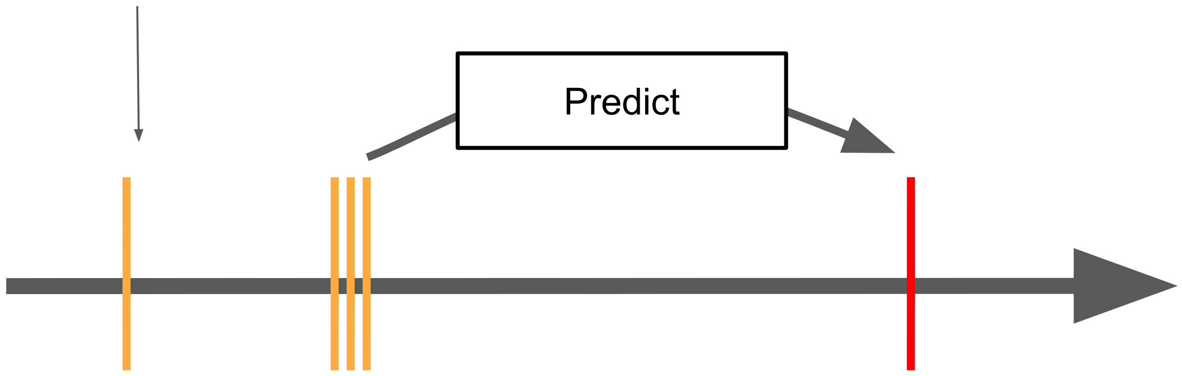
- Numerical values (Ex:Lab test, vital sign)
- One-hot encoding (Ex:Diagnostic, prescription, treatment)
- Free notes (Ex:Clinical notes, specialist report)
- Image (Ex:Scan)
- Genomics (Ex:23andMe)
- Wearable (Ex:Google Fit)

Electronic Health Records

Clinical note

Homme, 24 ans, électricien, douleur au coude droit, consomme drogue régulièrement

prescription: Aspirin 100 mg au besoin



- Numerical values (Ex:Lab test, vital sign)
- One-hot encoding (Ex:Diagnostic, prescription, treatment)
- Free notes (Ex:Clinical notes, specialist report)
- Image (Ex:Scan)
- Genomics (Ex:23andMe)
- Wearable (Ex:Google Fit)

Goals: Predict future disease

Applications:
- Clinical Support Decision
- Increase doctor efficiency
- Prevent medical errors
- Replace doctor (???)

Free text vs structured data

Fast and flexible

Homme, 24 ans, électricien, douleur au coude droit, consomme drogue régulièrement

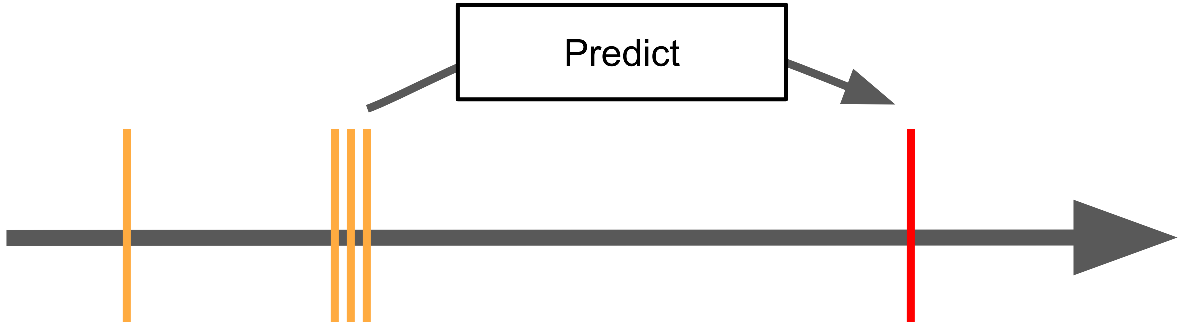
prescription: Aspirin 100 mg au besoin

"Hard" to use in ML

Slow and rigid

- Homme
- Age
- Aspirin Rx
- ? Alcohol

"Easy" to use in ML



Goals: Predict future disease

Applications:
- Clinical Support Decision
- Increase doctor efficiency
- Prevent medical errors
- Replace doctor (???)

Free text vs structured data

Fast and flexible

Homme, 24 ans, **électricien**, douleur au coude droit, consomme drogue régulièrement

prescription: Aspirin 100 mg au besoin

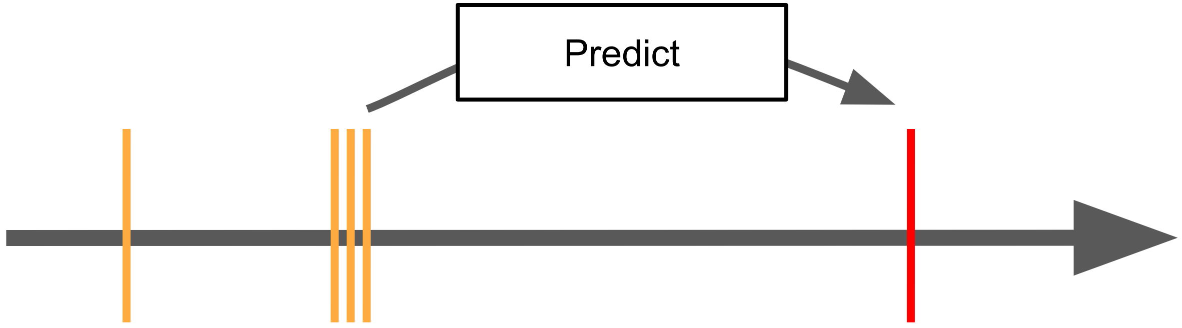
“Hard” to use in ML

Encode

Slow and rigid

- Homme
- Age
- Aspirin Rx
- ? Alcohol
- Électricien

“Easy” to use in ML



Goals: Predict future disease

Applications:
- Clinical Support Decision
- Increase doctor efficiency
- Prevent medical errors
- Replace doctor (???)

Free text vs structured data

Fast and flexible

Homme, 24 ans, électricien, **douleur** au **coude droit**, consomme drogue régulièrement

prescription: Aspirin 100 mg au besoin

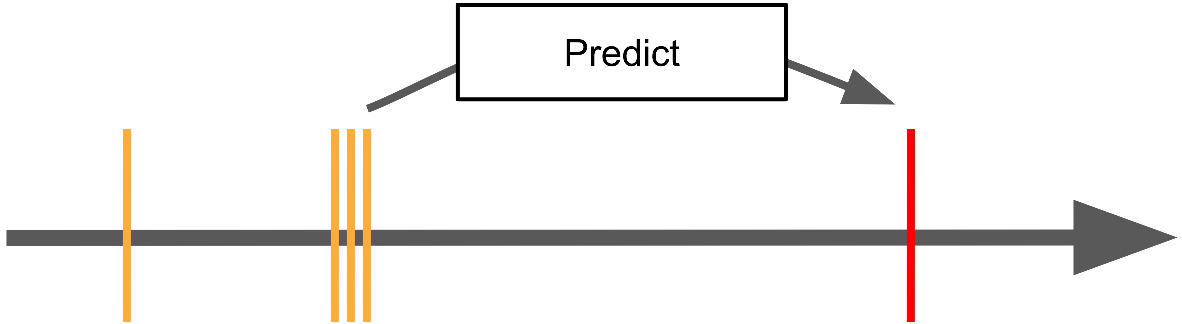
Encode

Slow and rigid

- Elbow
- Right Elbow
- Pain
- Pain Right Elbow
- Pain Elbow

"Hard" to use in ML

"Easy" to use in ML



Goals: Predict future disease

Applications:
- Clinical Support Decision
- Increase doctor efficiency
- Prevent medical errors
- Replace doctor (???)

Deep EHR

Deep EHR: Chronic Disease Prediction Using Medical Notes

Jingshu Liu* JINGSHU.LIU@NYU.EDU
New York University

Zachariah Zhang* ZZ1409@NYU.EDU
New York University

Narges Razavian NARGES.RAZAVIAN@NYUMC.ORG
New York University

Deep EHR

Deep EHR: Chronic Disease Prediction Using Medical Notes

Jingshu Liu* JINGSHU.LIU@NYU.EDU
New York University
Zachariah Zhang* ZZ1409@NYU.EDU
New York University
Narges Razavian NARGES.RAZAVIAN@NYUMC.ORG
New York University

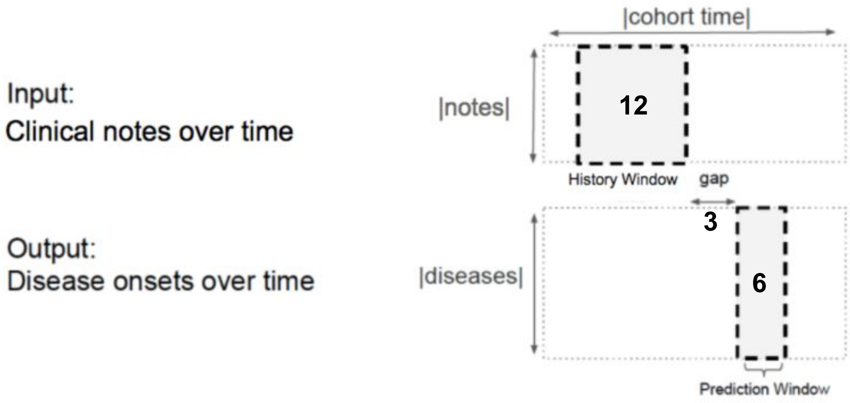


Figure 1: Overview of prediction framework

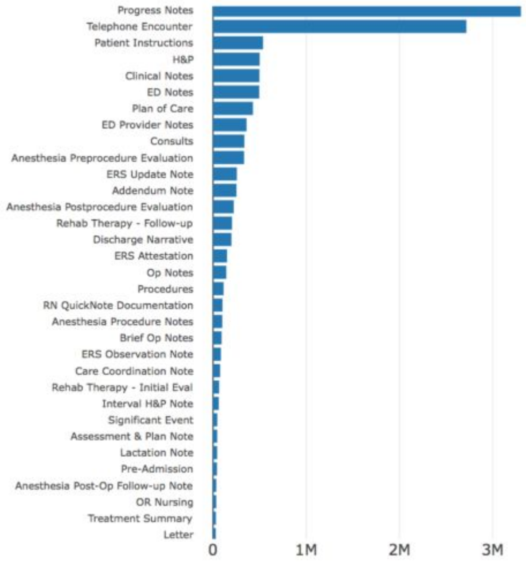


Figure 8: Note Type Distribution

We use medical notes, demographics and diagnoses in ICD-10 codes from the NYU Langone Hospital EHR system. The data contains clinical encounters of more than **1 million** patients between 2014 and 2017, and more than **15 million entries of medical notes**

Clinical notes and word processing

Notes pre-processing

- Kept 20k most frequent word
- Removed word > 80% prevalence
- Deidentified names, addresses, and locations with generic tokens

Notes stat

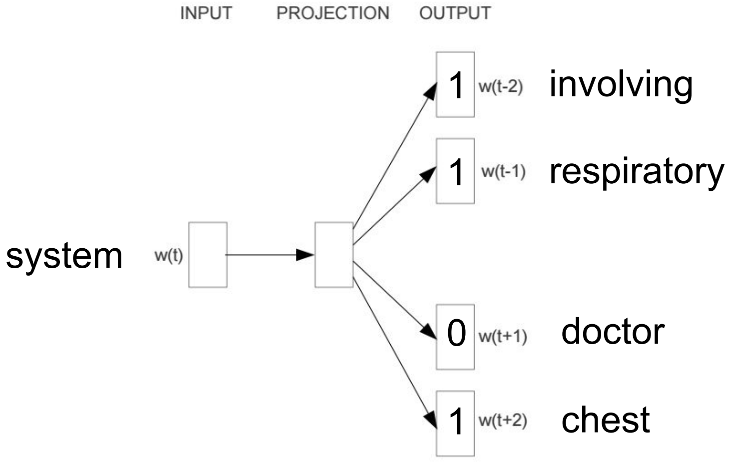
- Notes on average have 1300 words (90th percentile = 3000)

Word embedding

- Word2vec on Pubmed (23 millions documents, 24 millions unique word, 5.5 billions token)
- StarSpace (Ledell Wu *et al.* Starspace: Embed all the things!, 2017.)

word2vec

Context word	Context word	Target word		Context word	Context word	Context word
involving	respiratory	system	and	other	chest	symptoms



Skip-gram

Stolen from Joseph Paul Cohen

StarSpace: Embed All The Things!

Ledell Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes and Jason Weston
Facebook AI Research

In the general case, StarSpace embeds entities of **different types** into a **vectorial embedding space**, hence the “star” (“*”, meaning all types) and “space” in the name, and in that common space compares them against each other. It learns to rank a set of entities, documents or objects given a query entity, document or object, where the query is not necessarily of the same type as the items in the set.

$$\sum_{\substack{(a,b)\in E^+ \\ b^-\in E^-}} L^{batch}(\overbrace{sim(a,b)}^{\text{Positive pair}}, \overbrace{sim(a,b_1^-),\ldots,sim(a,b_k^-)}^{\text{Negative pair}})$$

Conclusions

- Text Classification / Sentiment Analysis: we show that our method achieves good results, comparable to fastText (Joulin et al. 2016) on three different datasets.
- Content-based Document recommendation: it can directly solve these tasks well, whereas applying off-the-shelf fastText, Tagspace or word2vec gives inferior results.
- Link Prediction in Knowledge Bases: we show that our method outperforms several methods, and matches TransE (Bordes et al. 2013) on Freebase 15K.
- Wikipedia Search and Sentence Matching tasks: it outperforms off-the-shelf embedding models due to directly training sentence and document-level embeddings.
- Learning Sentence Embeddings: It performs well on the 14 SentEval transfer tasks of (Conneau et al. 2017) compared to a host of embedding methods.

Clinical notes and word processing

Notes pre-processing

- Kept 20k most frequent word
- Removed word > 80% prevalence
- Deidentified names, addresses, and locations with generic tokens

Notes stat

- Notes on average have 1300 words (90th percentile = 3000)

Word embedding

- Word2vec on Pubmed (23 millions documents, 24 millions unique word, 5.5 billions token)
- StarSpace (Ledell Wu *et al.* Starspace: Embed all the things!, 2017.)

Extract lab values

- Use simple Regex to extract lab values from notes

Valx: A System for Extracting and Structuring Numeric Lab Test Comparison Statements from Text*

T. Hao[1,2]; H. Liu[3]; C. Weng[1]

[1]Department of Biomedical Informatics, Columbia University, New York, NY, USA;
[2]Key Lab of Language Engineering and Computing of Guangdong Province, Guangdong University of Foreign Studies, Guangzhou, China;
[3]Department of Health Sciences Research, Rochester, MN, USA

Table 1
The evaluation of Valx on Diabetes Type 2 and Type 1 diabetes trials using variable HbA1c compared with human-based reference standard dataset

Dataset	Text section	# by human	# by Valx	# Correct	Precision	Recall	F1
Diabetes Type 2	Inclusion	1934	1895	1877	99.1%	97.1%	98.0%
	Exclusion	186	184	177	96.2%	95.2%	95.7%
	Overall	2120	2079	2054	98.8%	96.9%	97.8%
Diabetes Type 1	Inclusion	403	397	396	99.7%	98.3%	99.0%
	Exclusion	66	65	64	98.5%	97.0%	97.7%
	Overall	469	462	460	99.6%	98.1%	98.8%
Both	Overall	2589	2541	2514	98.9%	97.1%	98.0%

Missing values are imputed with 0 if no previous test results exist for the same patient

Values of the top 50 most frequent lab tests are included in the model as

Table 3: Top 30 most frequent lab values

Item name	Prevalence by percentage of encounters
Weight	65.3%
BP_systolic	61.5%
BP_diastolic	61.4%
Height	55.2%
Pulse	54.7%
Oxygen saturation	38.2%
Temperature	37.9%
Resp	29.3%
BMI	27.7%
Urea Nitrogen	17.0%
Creatinine	16.7%
Chloride	15.3%
Potassium	15.1%
Sodium	15.0%
Carbon Dioxide	14.6%
Hemoglobin	14.0%
Hematocrit	13.3%
Glucose	13.3%
Alanine Aminotransferase	12.5%
Aspartate Aminotransferase	12.3%
Ery. Mean Corpuscular Volume	12.2%
Alkaline Phosphatase	10.7%
Bilirubin	10.6%
Platelets	10.5%
Calcium	10.3%
Leukocytes	5.7%
WBC	5.5%
HDL Cholesterol	5.3%
LDL Cholesterol	3.9%
Albumin	3.9%

Clinical notes and word processing

Notes pre-processing

- Kept 20k most frequent word
- Removed word > 80% prevalence
- Deidentified names, addresses, and locations with generic tokens

Notes stat

- Notes on average have 1300 words (90th percentile = 3000)

Word embedding

- Word2vec on Pubmed (23 millions documents, 24 millions unique word, 5.5 billions token)
- StarSpace (Ledell Wu *et al.* Starspace: Embed all the things!, 2017.)

Extract lab values

- Use simple Regex to extract lab values from notes

Negation tagging

- Negex system[Chapman et al. Chapman et al. (2001)]

Original note:

... no known allergies review of symptoms : general : no fevers , chills , or weight loss... no cough , shortness of breath , or wheezing cardiovascular : no chest pain or dyspnea on exertion gastrointestinal : no abdominal pain , change in bowel habits , or black or bloody stools... neurological : no transient ischemic attack or stroke symptoms...

Negation Tagged:

... no known **allergies_neg** review of symptoms : general : no **fevers_neg** , **chills_neg** , or **weight_neg loss_neg**... no **cough_neg** , **shortness_neg** of **breath_neg** , or **wheezing_neg** cardiovascular : no **chest_neg pain_neg** or **dyspnea_neg** on **exertion_neg** gastrointestinal : no **abdominal_neg pain_neg , change_neg in_neg bowel_neg habits_neg** , or black or bloody stools... neurological : no transient **ischemic_neg attack_neg** or **stroke_neg symptoms_neg**...

Clinical notes and word processing

Notes pre-processing
- Kept 20k most frequent word
- Removed word > 80% prevalence
- Deidentified names, addresses, and locations with generic tokens

Notes stat
- Notes on average have 1300 words (90th percentile = 3000)

Word embedding
- Word2vec on Pubmed (23 millions documents, 24 millions unique word, 5.5 billions token)
- StarSpace (Ledell Wu *et al.* Starspace: Embed all the things!, 2017.)

Extract lab values
- Use simple Regex to extract lab values from notes

Negation tagging
- Negex system[Chapman et al. Chapman et al. (2001)]

Demographic
- Age, sex, insurance type, etc.

Predictive task

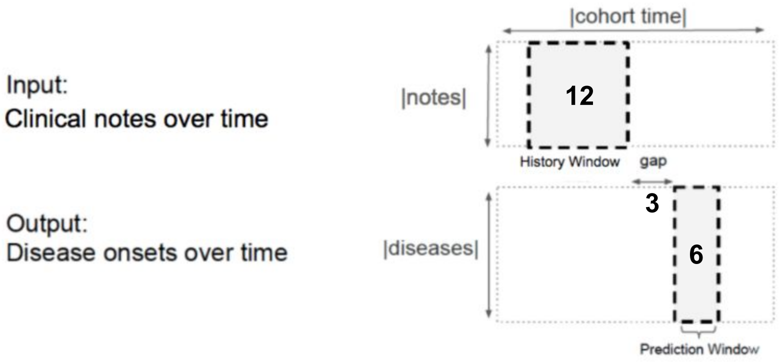


Figure 1: Overview of prediction framework

Table 1: Number of Records by Target Diseases (Negative Cases : Positive Cases)

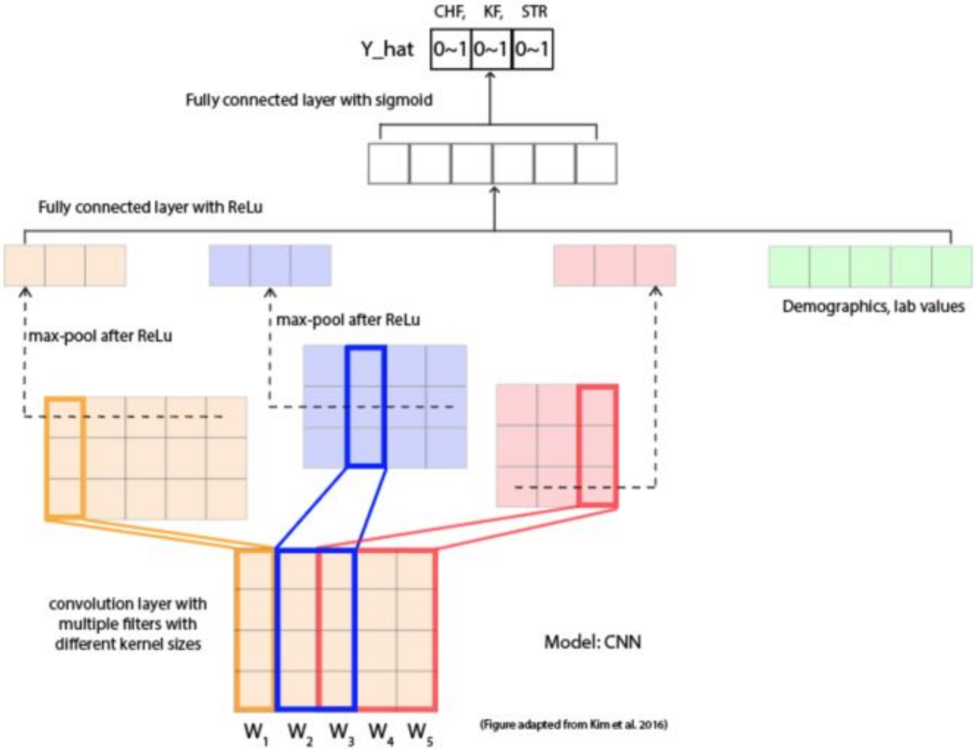
Target	Training Set	Validation Set	Test Set
Congestive Heart Failure	644K : 4080	93K : 574	184K : 1167
Kidney Failure	616K : 10051	88K : 1428	176K : 2809
Stroke	653K : 3195	94K : 406	187K : 916

Baseline

Table 2: Model Performance (AUC) by Target Disease

	Heart Failure	**Kidney Failure**	**Stroke**
Logistic Reg Lab/Demo	0.781	0.724	0.70
LSTM Lab/Demo	0.813	0.743	0.699
Logistic Reg Notes	0.810	0.752	0.708

CNN



CHF, KF, STR
Y_hat
0~1 0~1 0~1
Fully connected layer with sigmoid
Fully connected layer with ReLu
max-pool after ReLu
max-pool after ReLu
Demographics, lab values
convolution layer with multiple filters with different kernel sizes
Model: CNN
W1 W2 W3 W4 W5
(Figure adapted from Kim et al. 2016)

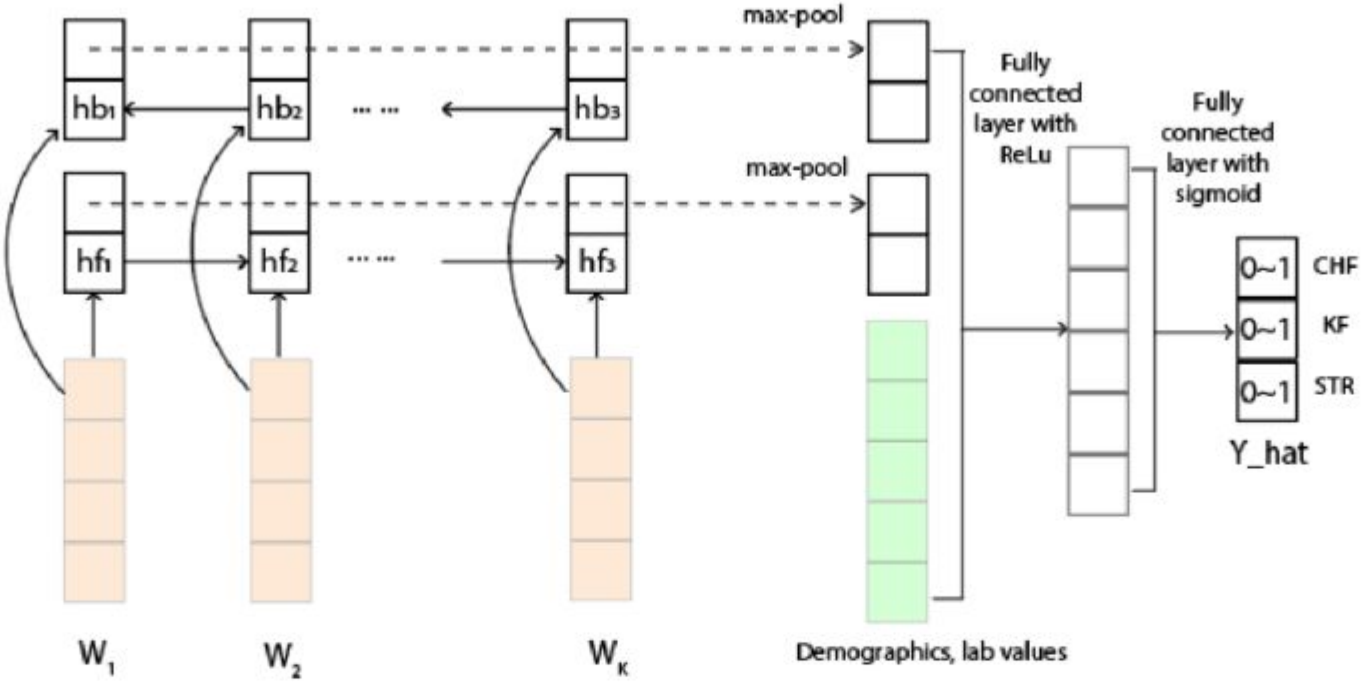
CNN

Table 2: Model Performance (AUC) by Target Disease

	Heart Failure	**Kidney Failure**	**Stroke**
Logistic Reg Lab/Demo	0.781	0.724	0.70
LSTM Lab/Demo	0.813	0.743	0.699
Logistic Reg Notes	0.810	0.752	0.708
CNN PubMed Embeddings	0.844	0.799	0.711
CNN Single Task	0.847	0.796	0.706
CNN	0.854	0.802	0.714
CNN + Neg Tag	0.867	0.811	0.727
CNN + Neg Tag + Dense	0.880	0.812	0.733
CNN + Neg Tag + Dense + Lab/Demo	0.893	0.822	0.749

(a) (b)

BiLSTM

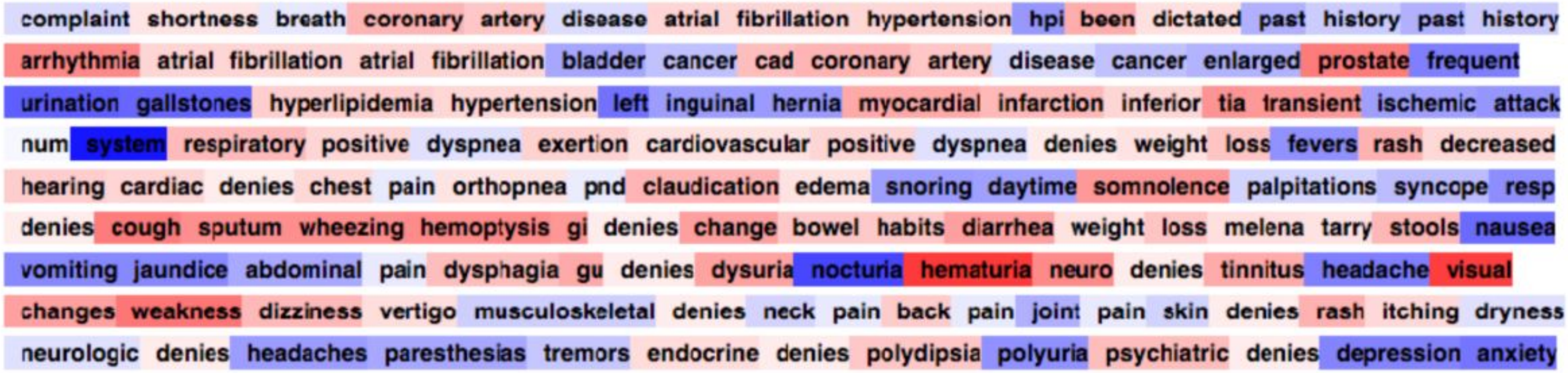

max-pool
max-pool
hb₁
hb₂
hb₃
hf₁
hf₂
hf₃
Fully connected layer with ReLu
Fully connected layer with sigmoid
0~1
0~1
0~1
CHF
KF
STR
Y_hat
W₁
W₂
W_K
Demographics, lab values

BiLSTM

Table 2: Model Performance (AUC) by Target Disease

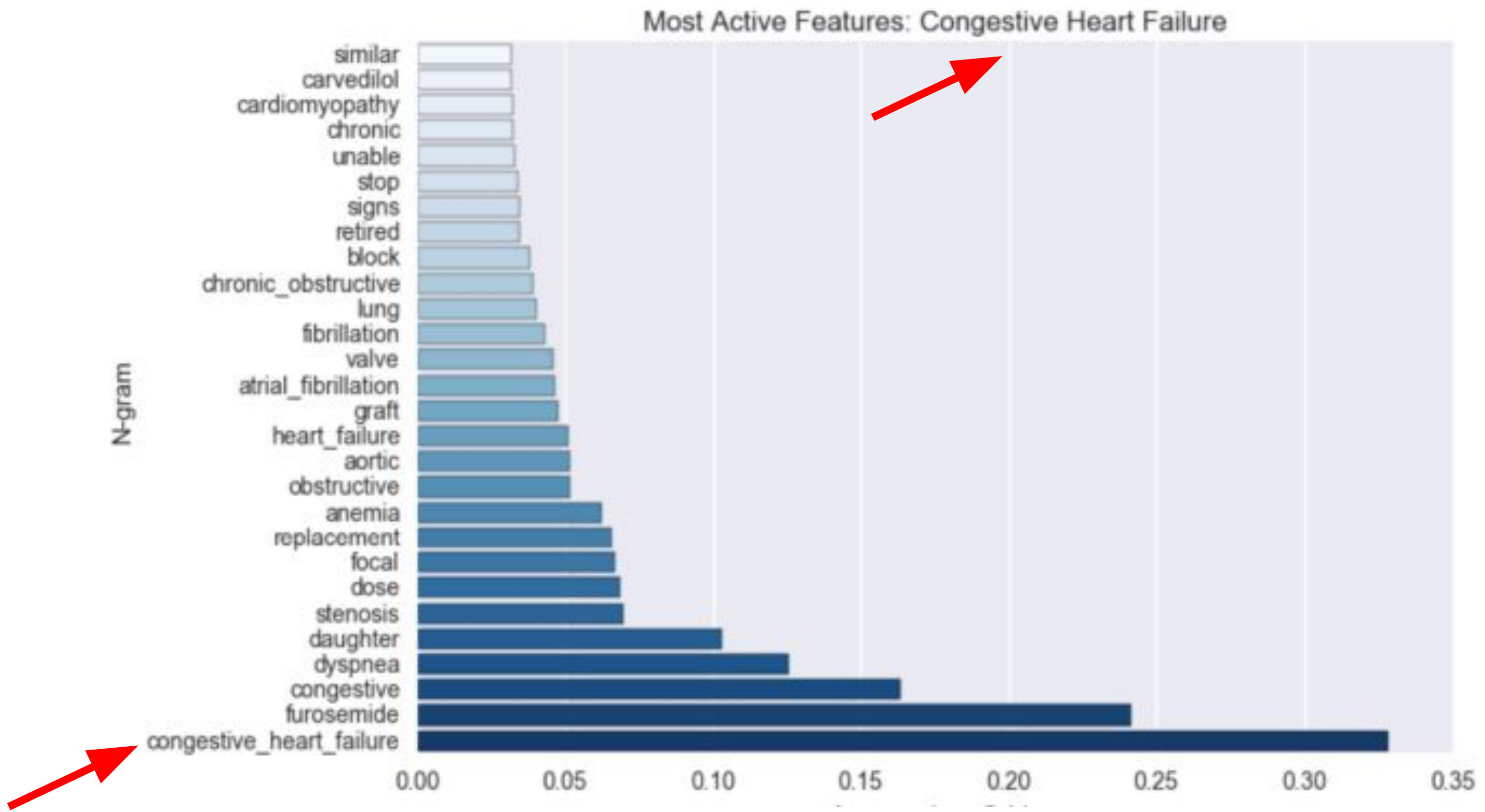
	Heart Failure	**Kidney Failure**	**Stroke**
Logistic Reg Lab/Demo	0.781	0.724	0.70
LSTM Lab/Demo	0.813	0.743	0.699
Logistic Reg Notes	0.810	0.752	0.708
CNN PubMed Embeddings	0.844	0.799	0.711
CNN Single Task	0.847	0.796	0.706
CNN	0.854	0.802	0.714
CNN + Neg Tag	0.867	0.811	0.727
CNN + Neg Tag + Dense	0.880	0.812	0.733
CNN + Neg Tag + Dense + Lab/Demo	0.893	0.822	0.749
BiLSTM	0.869	0.807	0.738
BiLSTM + Neg Tag	0.875	0.811	0.745
BiLSTM + Neg Tag + Dense	0.892	0.823	0.739
BiLSTM + Neg Tag + Dense + Lab/Demo	**0.900**	**0.833**	**0.753**

Explanation

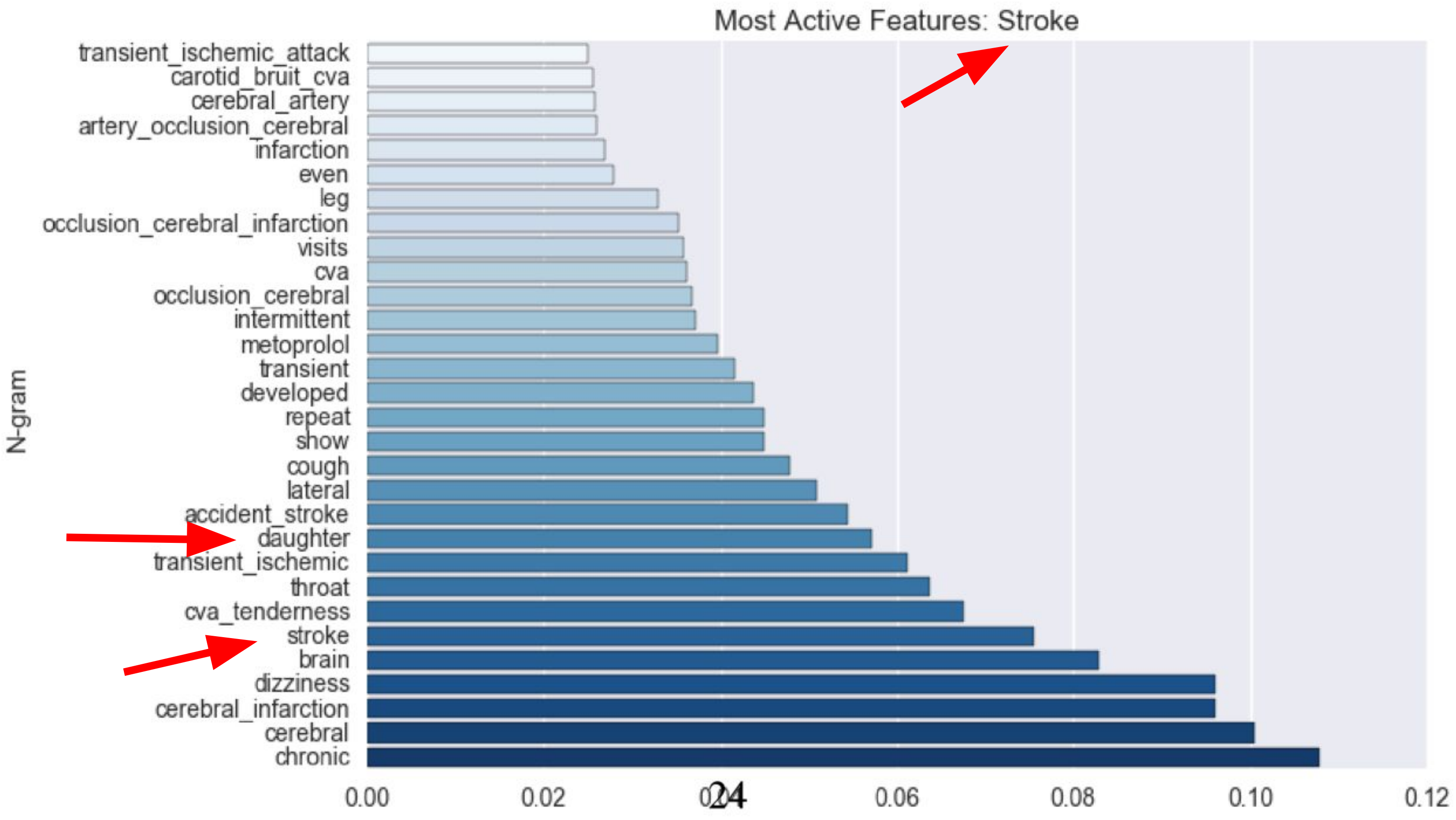


Contribution to prediction of heart failure prediction

Bias in the data



Bias in the data



Lessons from Natural Language Inference in the Clinical Domain

Alexey Romanov
Department of Computer Science
University of Massachusetts Lowell*
Lowell, MA 01854
aromanov@cs.uml.edu

Chaitanya Shivade
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120
cshivade@us.ibm.com

Goals and Tasks

Natural language inference (**NLI**) is the task of determining whether a given hypothesis can be inferred from a given premise.

"We have presented MedNLI, an expert annotated, public dataset for **natural language inference in the clinical domain**."

#	Premise	Hypothesis	Label
1	ALT , AST , and lactate were elevated as noted above	patient has abnormal lfts	entailment
2	Chest x-ray showed mild congestive heart failure	The patient complains of cough	neutral
3	During hospitalization , patient became progressively more dyspnic requiring BiPAP and then a NRB	The patient is on room air	contradiction
4	She was not able to speak , but appeared to comprehend well	Patient had aphasia	entailment
5	T1DM : x 7yrs , h/o DKA x 6 attributed to poor medication compliance , last A1c [** 3-23 **] : 13.3 % 2	The patient maintains strict glucose control	contradiction
6	Had an ultimately negative esophagogastroduodenoscopy and colonoscopy	Patient has no pain	neutral
7	Aorta is mildly tortuous and calcified .	the aorta is normal	contradiction

Table 1: Examples from the development set of MedNLI

Data

- “As the source of premise sentences, we used the MIMIC-III v1.3 (Johnson et al., 2016) database.”
- Ask clinician this task ->

You will be shown a sentence from the `Past Medical History` section of a de-identified clinical note. Using only this sentence, your knowledge about the field of medicine, and common sense:

- Write one alternate sentence that is **definitely** a **true** description of the patient. Example, for the sentence “Patient has type II diabetes” you could write “Patient suffers from a chronic condition“
- Write one alternate sentence that **might be** a **true** description of the patient. Example, for the sentence “Patient has type II diabetes” you could write “Patient has hypertension”
- Write one sentence that is **definitely** a **false** description of the patient. Example, for the sentence “Patient has type II diabetes” you could write “The patient’s insulin levels are normal without any medications.”

Data

- “As the source of premise sentences, we used the MIMIC-III v1.3 (Johnson et al., 2016) database.”
- Ask clinician this task ->
  - 2 radiologist (100 premises each)
  - 552 pairs of premises/hypothesis
  - Reviewed by independent clinician
    - Cohen’s kappa κ = 0.78
  - Recruited 2 residents
    - 4,683 premises over a period of six weeks
    - 14,049 unique sentence pairs

Dataset size	
Training pairs	11232
Development pairs	1395
Test pairs	1422
Average sentence length in tokens	
Premise	20.0
Hypothesis	5.8
Maximum sentence length in tokens	
Premise	202
Hypothesis	20

Table 2: Key statistics of the dataset

Models

Bag-of-words
“In order to represent an input sentence as a single vector, this architecture simply sums up the vectors of individual tokens. The premise and hypothesis vectors are then concatenated and passed through a multi-layer neural network.”

Models

Bag-of-words
"In order to represent an input sentence as a single vector, this architecture simply sums up the vectors of individual tokens. The premise and hypothesis vectors are then concatenated and passed through a multi-layer neural network."

Infersent
"InferSent (Conneau et al., 2017) is a model for sentence representation that demonstrated close to state-of-the-art performance across a number of tasks in NLP (including NLI) and computer vision."

"A bidirectional LSTM encoder of input sentences and a max-pooling operation over timesteps are used to get a vector for the premise (p) and for the hypothesis (h)"

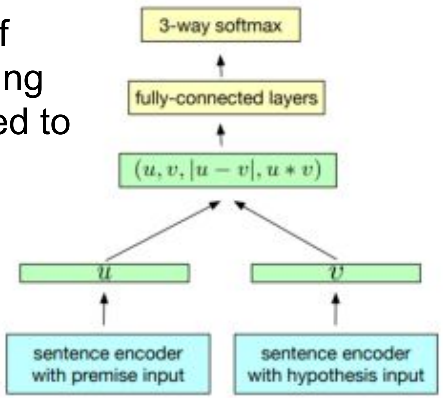


Figure 1: **Generic NLI training scheme**

Supervised Learning of Universal Sentence Representations from Natural Language Inference Data

Alexis Conneau Facebook AI Research aconneau@fb.com

Douwe Kiela Facebook AI Research dkiela@fb.com

Holger Schwenk Facebook AI Research schwenk@fb.com

Loïc Barrault LIUM, Université Le Mans loic.barrault@univ-lemans.fr

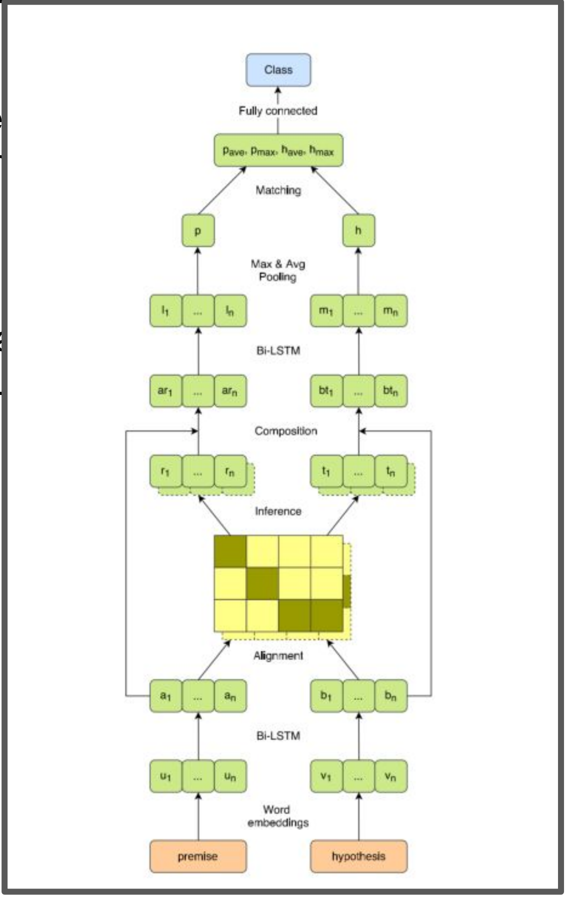
Antoine Bordes Facebook AI Research abordes@fb.com

Models

Bag-of-words
“In order to represent an input sentence as a single vector, this architecture
of individual tokens. The premise and hypothesis vectors are then concaten
multi-layer neural network.”

Infersent
“InferSent (Conneau et al., 2017) is a model for sentence representation tha
state-of-the-art performance across a number of tasks in NLP (including NL

ESIM
“It is a fairly complex model that makes use of two bidirectional
LSTM networks.”



Models

Bag-of-words

"In order to represent an input sentence as a single vector, this architecture simply sums up the vectors of individual tokens. The premise and hypothesis vectors are then concatenated and passed through a multi-layer neural network."

Infersent

"InferSent (Conneau et al., 2017) is a model for sentence representation that demonstrated close to state-of-the-art performance across a number of tasks in NLP (including NLI) and computer vision."

ESIM

"It is a fairly complex model that makes use of two bidirectional LSTM networks."

Feature-Based system

"gradient boosting classifier incorporating"

The groups below summarize the feature sets used in our model (35 features in total):

1. BLEU score
2. Number of tokens (e.g. min, max, difference)
3. Negations (e.g. keywords such as *no, do not*)
4. TF-IDF similarity (e.g. cosine, euclidean)
5. Edit distances (e.g. Levenshtein)
6. Embedding similarity (e.g. cosine, euclidean)
7. UMLS similarity features (e.g. shortest path distance between UMLS concepts)

Baseline

Set	Features	BOW	InferSent	ESIM
Dev	51.9	71.9	**76.0**	74.4
Test	51.9	70.2	**73.5**	73.1

Table 4: Baseline accuracy on the development and the test set of MedNLI for different models.

Transfert learning

Source domain	Direct transfer			Sequential transfer			Multi-target transfer		
	BOW	InferSent	ESIM	BOW	InferSent	ESIM	BOW	InferSent	ESIM
snli	-21.8	**-24.2**	-22.8	1.8	-1.8	-2.5	**2.4**	-2.5	-0.7
fiction	**-21.6**	-25.6	**-21.4**	1.3	0.4	-0.5	1.4	0.1	**0.3**
government	-23.8	-27.2	-26.2	1.0	0.8	-0.7	1.3	0.2	0.2
slate	-23.2	-25.7	-21.6	**1.9**	**0.9**	**-0.2**	1.1	**0.6**	-0.1
telephone	-25.7	-27.3	-25.6	1.7	-0.2	-1.1	1.2	0.4	-0.1
travel	-25.4	-29.1	-23.5	1.6	0.0	-0.7	0.2	-0.3	0.1

Table 5: Absolute gain in accuracy with respect to the baseline (see Table 4) on the MedNLI test set for different transfer learning modes. Bold indicates the best source domain for each model and transfer.

Direct transfer: Trained on a source domain and test on MedNLI
Sequential transfert: “After pre-training the model on a large source domain, the model is further fine-tuned us- ing the smaller training data of the target domain.”

Multi-target transfert learning

- **The shared component** is trained on both the source and target domains
- **The source domain component** is trained only during the pre-training phase and does not participate in the prediction of the target domain
- **The target domain component** is trained during the fine-tuning stage and it produces the predictions together with the shared component.

The motivation for multi-target transfer is that the performance should improve by splitting deeper layers of the model into domain-specific parts and having a shared block early in the network, **where it presumably learns domain independent features**. The target-specific component will not be in the local minimum of the source domain after the pre-training stage, enabling the model to find a better local minimum for the target domain.

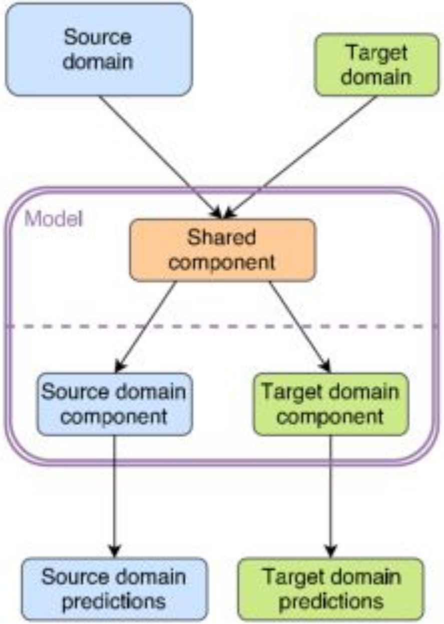


Figure 4: Schematic depiction of the model for multi-target transfer learning

Word Embedding

Embeddings	BOW	InferSent	ESIM
$\text{fastText}_{[\text{Wiki}]}$	-3.5	-3.5	-4.4
$\text{fastText}_{[\text{CC}]}$ (600B)	-0.6	1.3	-0.3
$\text{fastText}_{[\text{BioASQ}]}$ (2.3B)	0.5	0.6	0.2
$\text{fastText}_{[\text{MIMIC-III}]}$ (0.8B)	**1.1**	2.3	1.2
$\text{GloVe}_{[\text{CC}]} \rightarrow \text{fastText}_{[\text{BioASQ}]}$	0.2	0.7	1.4
$\text{GloVe}_{[\text{CC}]} \rightarrow \text{fastText}_{[\text{BioASQ}]} \rightarrow \text{fastText}_{[\text{MIMIC-III}]}$	0.9	2.7	**1.8**
$\text{fastText}_{[\text{Wiki}]} \rightarrow \text{fastText}_{[\text{MIMIC-III}]}$	0.1	**3.1**	1.7

Wiki -> Wikipedia
CC -> Common Crawl
BioASQ -> Pubmed
MIMIC-III -> Clinical Notes

Knowledge Graph - UMLS - SNOMED T

“SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world. It is designed for use in **clinical documentation in the Electronic Health Record (EHR)**. The purpose of the SNOMED CT to ICD-10-CM map (herein referred to as "the Map") is to support semi-automated generation of ICD-10-CM codes from clinical data encoded in SNOMED CT for reimbursement and statistical purposes.”

Knowledge Graph - UMLS - SNOMED T

“SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world. It is designed for use in **clinical documentation in the Electronic Health Record (EHR)**. The purpose of the SNOMED CT to ICD-10-CM map (herein referred to as "the Map") is to support semi-automated generation of ICD-10-CM codes from clinical data encoded in SNOMED CT for reimbursement and statistical purposes.”

- Standardize Clinical Notes (abbreviation)
- Map word to concept

Semantic type	**Common examples**	**Count**
Finding	*asymptomatic, history of kidney stones, nystagmus*	35,439
Disease or Syndrome	*chf, enterovaginal fistula, diverticulitis, acute stroke*	9,941
Sign or Symptom	*chest pain, dyspnea, seizures, vomiting, nausea*	5,294
Therapeutic Procedure	*aspiration, cabg, limb perfusion, chemotherapy*	5,043
Pharmacological Substance	*lopressor, morphine, atenolol, ativan, coumadin*	3,948
Body part, organ	*r arm, jaw, left frontal lobe brain, patellar tendon*	3,907
Laboratory Procedure	*serum glucose, blood ph, cbc, hematocrit, neutrophil count*	1,136

Table 3: Examples of medical concepts belonging to common semantic types across premises and hypotheses in the MedNLI training data.

Knowledge Graph - UMLS - SNOMED T

“SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world. It is designed for use in **clinical documentation in the Electronic Health Record (EHR)**. The purpose of the SNOMED CT to ICD-10-CM map (herein referred to as "the Map") is to support semi-automated generation of ICD-10-CM codes from clinical data encoded in SNOMED CT for reimbursement and statistical purposes.”

- Standardize Clinical Notes (abbreviation)
- Map word to concept
- Relation between words
  - 327,001 entities
  - 3,809,639 edges
  - 169 different types of edge

Each terminology in the UMLS can be viewed as a graph where nodes represent medical concepts, and edges represent relations between them. These are canonical relationships found in ontologies such as IS A and SYNONYMY.

For instance, diabetes IS A disorder of the endocrine system.

Knowledge Graph - UMLS - SNOMED T

“SNOMED CT is considered to be the most comprehensive, multilingual clinical healthcare terminology in the world. It is designed for use in **clinical documentation in the Electronic Health Record (EHR)**. The purpose of the SNOMED CT to ICD-10-CM map (herein referred to as "the Map") is to support semi-automated generation of ICD-10-CM codes from clinical data encoded in SNOMED CT for reimbursement and statistical purposes.”

- Standardize Clinical Notes (abbreviation)

- Map word to concept

- Relation between words
  - 327,001 entities
  - 3,809,639 edges
  - 169 different types of edge

Knowledge Integration
- Retrofitting: Word embedding from a graph -> Connected word should be close in the embedded space
- Knowledge-directed attention

Retrofitting

BOW	InferSent	ESIM
-1.7	-2.0	-2.7

Table 7: Absolute gain in accuracy using retrofitting for MedNLI.

Retrofitting is "agnostic" of the edge type

Retrofitting is good for synonym, but connected medical concept in SNOMED CT are not necessarily synonym

Distance between premise and hypothesis is often 0 or no path

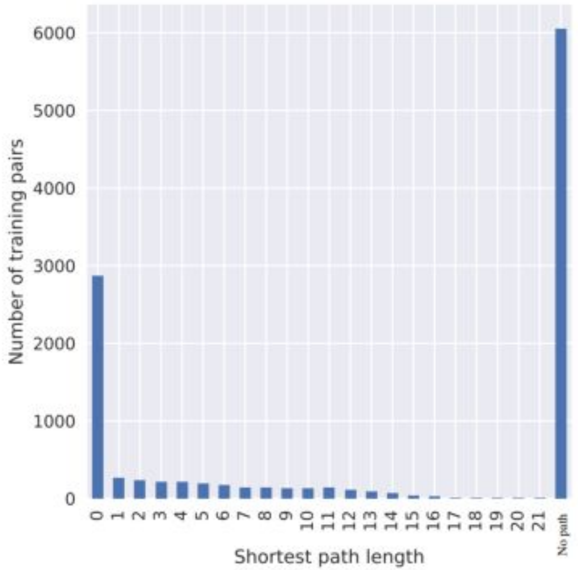


Figure 5: Lengths of the shortest paths between concepts in the premise and the hypothesis. 0 indicates that they contain the same concept.

Knowledge-directed attention

“We propose to integrate this knowledge in a way similar to how attention is used in the ESIM model. Specifically, we calculate the **attention matrix** e ∈ R n×m between **all pairs of tokens ai and bj** in the inputs sentences, where n is the length of the hypothesis and m is the length of the premise. **The value in each cell reflects the length of the shortest path lij** between the corresponding concepts of the premise and the hypothesis in SNOMED-CT.”

For example, there is an edge in SNOMED-CT from the concept *Lung consolidation* to *Pneumonia*. Using this information, during the processing of a sentence pair

- **Premise** The patient has *pneumonia*.
- **Hypothesis** The patient has a *lung* disease.

the model could attend to the token *lung* while processing *pneumonia*.

Embedding	**InferSent**	**ESIM**
$\text{GloVe}_{[\text{CC}]}$	0.3	0.0
$\text{fastText}_{[\text{MIMIC-III}]}$	0.2	0.3

Table 8: Absolute gain in accuracy using knowledge-directed attention.

“Interestingly, gains from **knowledge-directed attention stem mostly (60%) from the neutral class**. Moreover, 87% of these neutral predictions were predicted as entailment before adding the knowledge directed attention.”

Where does it fail?

Category	Premise	Hypothesis	Predicted	Expected
Numerical	On weaning to 6LNC, his O2 decreased to 81-82%	He has poor O2 stats	neutral	entailment
Reasoning	WBC 12 , Hct 41 .	WBC slightly elevated	contradiction	entailment
World	The infant emerged with spontaneous cry.	The infant was still born.	entailment	contradiction
Knowledge	No known sick contacts	No recent travel	entailment	neutral
Abbreviation	No CP or fevers.	Patient has no angina	neutral	entailment
	Received GI cocktail for h/o GERD, esophageal spasm	Received a proton pump inhibitor	entailment	neutral
Medical	EKG showed T-wave depression in V3-5, with no prior EKG for comparison.	Patient has a normal EKG	neutral	contradiction
Knowledge	Mother developed separation of symphysis pubis and was put in traction .	She has orthopedic injuries	neutral	entailment
Negation	Head CT was negative for bleed.	The patient has intracranial hemorrhage	neutral	contradiction
	Denied headache, sinus tenderness, or congestion	Patient has headaches	neutral	contradiction

Bias in the dataset

#	Premise	Hypothesis	Label
1	ALT , AST , and lactate were elevated as noted above	patient has abnormal lfts	entailment
2	Chest x-ray showed mild congestive heart failure	The patient complains of cough	neutral
3	During hospitalization , patient became progressively more dyspnic requiring BiPAP and then a NRB	The patient is on room air	contradiction
4	She was not able to speak , but appeared to comprehend well	Patient had aphasia	entailment
5	T1DM : x 7yrs , h/o DKA x 6 attributed to poor medication compliance , last A1c [** 3-23 **] : 13.3 % 2	The patient maintains strict glucose control	contradiction
6	Had an ultimately negative esophagogastroduodenoscopy and colonoscopy	Patient has no pain	neutral
7	Aorta is mildly tortuous and calcified .	the aorta is normal	contradiction

Table 1: Examples from the development set of MedNLI

Predict

F1 Score = 61.9
(SNLi F1 Score = 67.0)

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. Proceedings of NAACL.

Mila Medical

Joseph Paul Cohen, PhD

Pr. Yoshua Bengio

Francis Dutil

Martin Weiss

Shawn Tan

Tristan Sylvain

Margaux Luck, PhD

Assya Trofimov

Vincent Frappier, PhD