TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Pretraining for NLP

Pretraining for NLP

In NLP unsupervised pretraining is now required for strong benchmark performance.

Pretrained Word Embeddings

Advances in Pre-Training Distributed Word Representations, Mikolov et al., 2017

We want a mapping from a word w to a vector $e(w)$ — a word embedding.

fastText from Facebook is currently popular.

It provides both contextual bag of words (cbow) and byte pair encoding (BPE) word vectors.

cbow word vectors

We construct a population distribution on pairs (c, w) here c is a bag of word context and w is a word.

$$\Phi^* = \operatorname*{argmin}_{\Phi} E_{c,w} \; - \ln P(w|c)$$

Φ consists of a matrix $e[w, i]$ where $e[w, I]$ is the word embedding of w, and a matrix $e'[w, i]$ giving the embedding of the word w when it appears in a context.

A score $s(w|c)$ is defined by

$$s(w|c) = \frac{1}{|c|} \sum_{w' \in c} \; e(w)^\top e'(w')$$

Negative Sampling in cbow

Rather than define $P_\Phi(w|c)$ by a softmax over w, one uses restricted negative sampling.

We construct a training set of triples (w, c, N_C)

$$\Phi^* = \operatorname*{argmin}_{\Phi} \; E_{w,c,N_c} \; \ln\left(1 + e^{-s(w,c)}\right) + \sum_{n \in N_C} \ln\; \left(1 + e^{s(n,c)}\right)$$

Byte Pair Encoding (BPE)

BPE constructs a set of character n-grams by starting with the unigrams and then greedily merging most common bigrams of n-grams.

Given a set of character n-grams each word is treated as a bag of character n-grams.

$$e[w] = \frac{1}{N} \sum_{n \in w} e(n)$$

Current systems use byte pairs but train the byte pair embeddings as part of transformer training.

BERT: Blank Languagage Modeling

We replace a random subset of the words with a blank token.

We run a transformer on a block of text containing some blanks.

For a blank occurring at position t we predict the word at position t:

$$P(w) = \operatorname*{softmax}_{w}\ h[t, J]e[w, J]$$

Blank language modeling outperforms language modeling when used for pretraining in classification tasks such as the GLUE tasks.

GLUE

GLUE: General Language Understanding Evaluation

ArXiv 1804.07461

Corpus	|Train|	|Test|	Task	Metrics	Domain
Single-Sentence Tasks					
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
Inference Tasks					
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

GLUE Leader Board as of February 27, 2020

	Rank	Name	Model	URL	Score
	1	T5 Team - Google	T5		90.3
	2	ERNIE Team - Baidu	ERNIE		90.1
	3	Microsoft D365 AI & MSR AI & GATECH	MT-DNN-SMART		89.9
+	4	王玮	ALICE v2 large ensemble (Alibaba DAMO NLP)		89.7
+	5	Microsoft D365 AI & UMD	FreeLB-RoBERTa (ensemble)		88.4
	6	Junjie Yang	HIRE-RoBERTa		88.3
	7	Facebook AI	RoBERTa		88.1
+	8	Microsoft D365 AI & MSR AI	MT-DNN-ensemble		87.6
	9	GLUE Human Baselines	GLUE Human Baselines		87.1

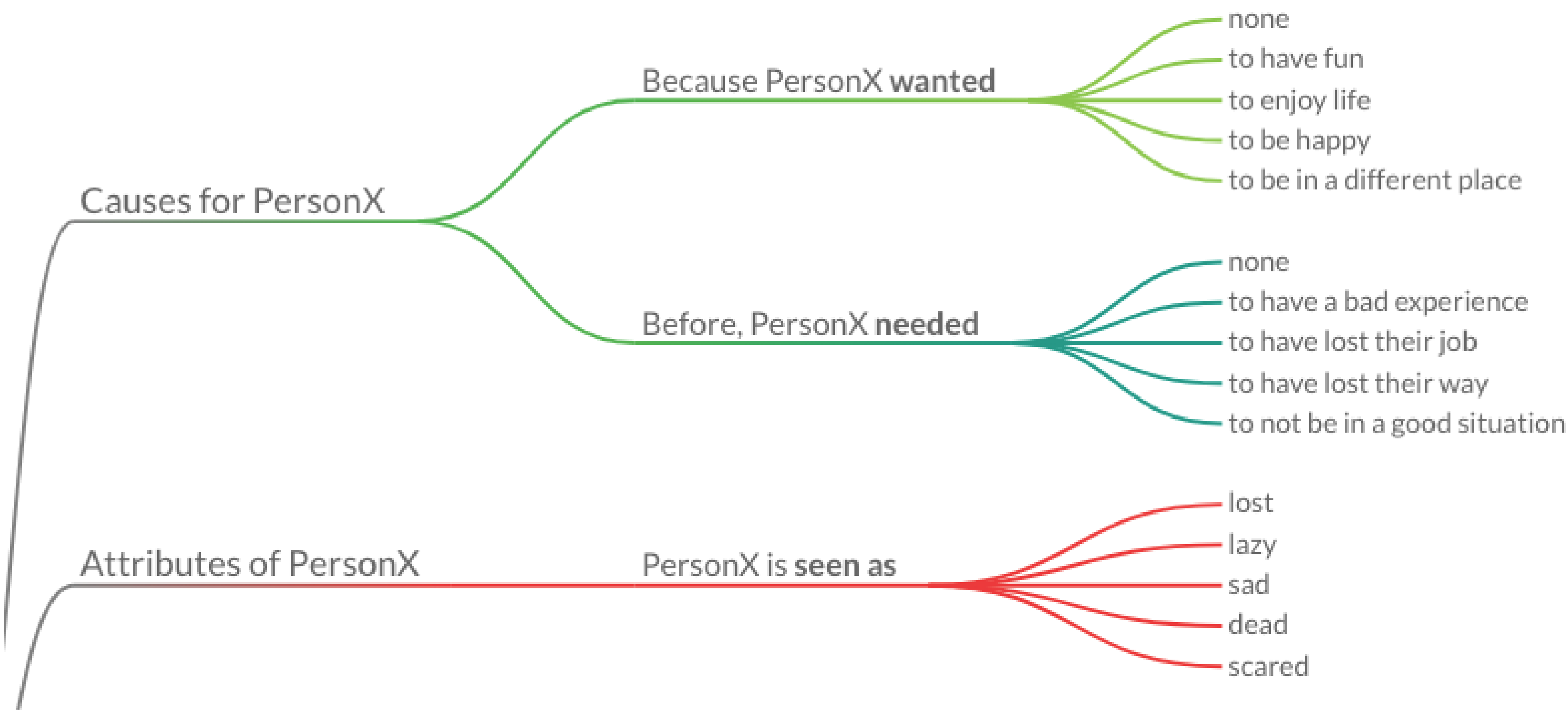
SuperGLUE Leader Board as of February 27, 2020

	Rank	Name	Model	URL	Score
	1	SuperGLUE Human Baselines	SuperGLUE Human Baselines		89.8
+	2	T5 Team - Google	T5		89.3
	3	Zhuiyi Technology	RoBERTa-mtl-adv		85.7
	4	Facebook AI	RoBERTa		84.6
	5	IBM Research AI	BERT-mtl		73.5

Fine Tuning on Question Answering

COMET: Busselut et al, June 2019.

Charlie is drifting though life:



The Chatbot Meena

User: *How are things?*

Meena: They're good. How about you?

User: *I'm trying to decide what to watch*

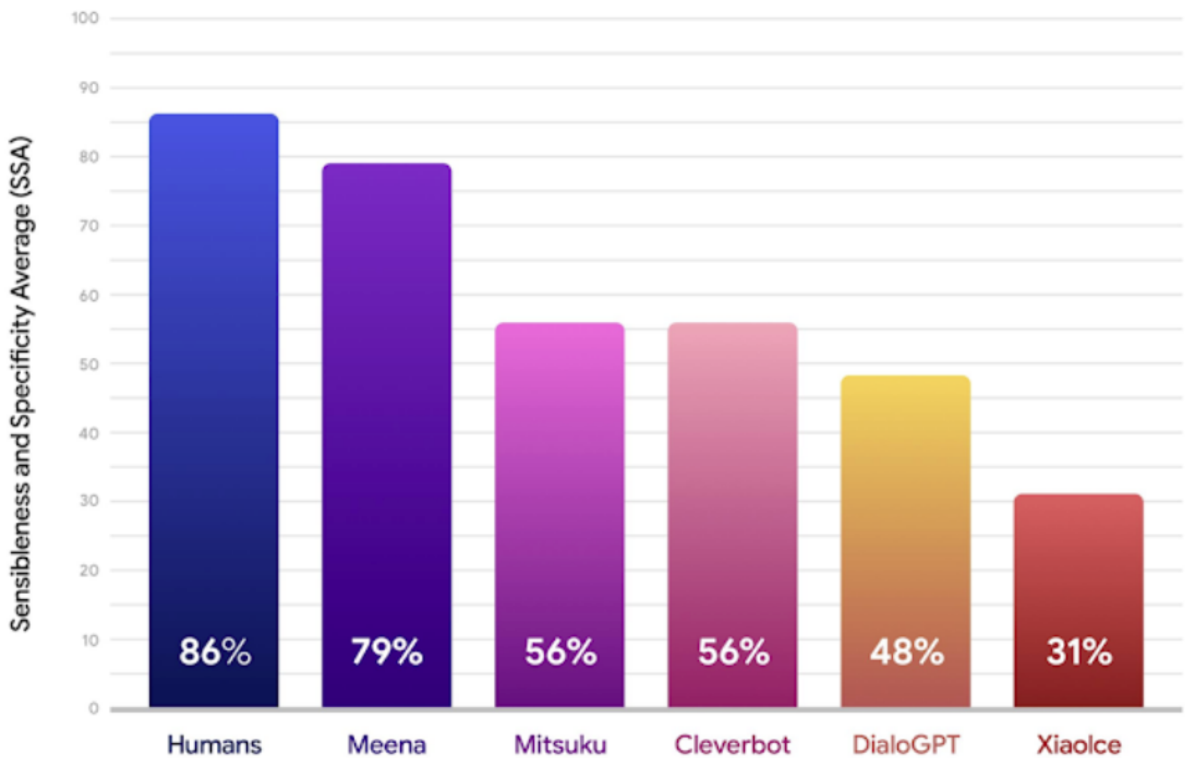
Meena: What are some of your favorite shows?

User: *I like the Good Place*

Meena: I haven't watched that. My favorite shows are Supernatural, and Star Trek: TNG.

User: *What is TNG?*

The Chatbot Meena



END