Natural Language Processing with Deep Learning

CS224N/Ling284

Christopher Manning

Lecture 9: Final Projects: Practical Tips

Lecture Plan

Lecture 9: Final Projects – practical tips – A pause for breath!

1. Final project types and details; assessment revisited
2. Finding research topics; a couple of examples
3. Finding data
4. Doing your research
5. Presenting your results and evaluation

1. Course work and grading policy

- 5 x 1-week Assignments: 6% + 4 x 12%: 54%
- Final Default or Custom Course Project (1–3 people): 43%
  - Project proposal: 5%; milestone: 5%; poster: 3%; report: 30%
  - Final poster session attendance expected! (See website.)
    Mon Mar 16, 5pm-10pm (put it in your calendar!)
- Participation: 3%
  - Guest/random lecture attendance, Piazza, eval, karma – see website!
    - Paul Butler for Piazza post explaining how to use Jupyter on Azure!
- Late day policy
  - 6 free late days; then 10% off per day; max 3 late days per assignment
- Collaboration policy: Read the website and the Honor Code!
  - For projects: It's okay to use existing code/resources, but you **must document** it, and you will be graded on your value-add
  - If multi-person: Include a brief statement on the work of each team-mate

Mid-quarter feedback survey

- Is out
- Please fill it in!
- We’d love to get your thoughts on the course so far!
- A good chance to improve the course immediately, as well as helping for future years
- Bribe: 0.5% participation points – make sure to submit the second form that records your name disassociated from the survey

The Final Project

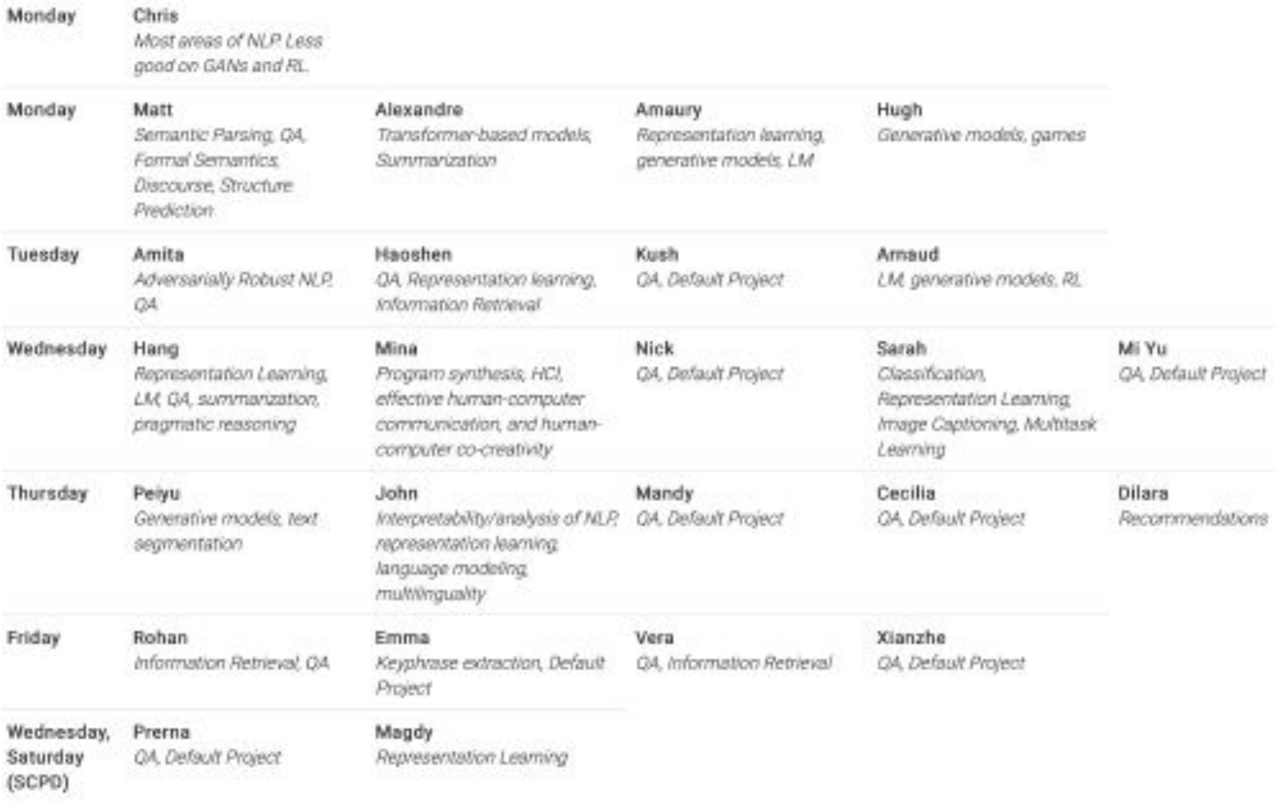
- For FP, you either
  - Do the default project, which is SQuAD question answering
    - Open-ended but an easier start; a good choice for most
  - Propose a custom final project, which we must approve
    - You will receive feedback from a **mentor** (TA/prof/postdoc/PhD)

- You can work in teams of 1–3
  - Larger team project or a project for multiple classes should be larger and often involve exploring more tasks

- You can use any language/framework for your project
  - Though we short of expect most of you to keep using PyTorch
  - And our starter code for the default FP is in PyTorch

Custom Final Project

- I’m very happy to talk to people about final projects, but the slight problem is that there’s only one of me….
- Look at TA expertise for custom final projects:
  - http://web.stanford.edu/class/cs224n/office_hours.html#staff



Monday	**Chris** *Most areas of NLP. Less good on GANs and RL.*				
Monday	**Matt** *Semantic Parsing, QA, Formal Semantics, Discourse, Structure Prediction*	**Alexandre** *Transformer-based models, Summarization*	**Amaury** *Representation learning, generative models, LM*	**Hugh** *Generative models, games*	
Tuesday	**Amita** *Adversarially Robust NLP, QA*	**Haoshen** *QA, Representation learning, Information Retrieval*	**Kush** *QA, Default Project*	**Arnaud** *LM, generative models, RL*	
Wednesday	**Hang** *Representation Learning, LM, QA, summarization, pragmatic reasoning*	**Mina** *Program synthesis, HCI, effective human-computer communication, and human-computer co-creativity*	**Nick** *QA, Default Project*	**Sarah** *Classification, Representation Learning, Image Captioning, Multitask Learning*	**Mi Yu** *QA, Default Project*
Thursday	**Peiyu** *Generative models, text segmentation*	**John** *Interpretability/analysis of NLP, representation learning, language modeling, multilinguality*	**Mandy** *QA, Default Project*	**Cecilia** *QA, Default Project*	**Dilara** *Recommendations*
Friday	**Rohan** *Information Retrieval, QA*	**Emma** *Keyphrase extraction, Default Project*	**Vera** *QA, Information Retrieval*	**Xianzhe** *QA, Default Project*	
Wednesday, Saturday (SCPD)	**Prerna** *QA, Default Project*	**Magdy** *Representation Learning*			

The Default Final Project

- There's a long handout on the web about it now!
- Task: Building a textual question answering system for SQuAD
  - Stanford Question Answering Dataset
    - https://rajpurkar.github.io/SQuAD-explorer/
  - Providing starter code in PyTorch ☺
  - Attempting SQuAD 2.0 (has unanswerable Qs)

- We will discuss question answering and SQuAD later. Example:

  T: [Bill] Aken, adopted by Mexican movie actress Lupe Mayorga, grew up in the neighboring town of Madera and his song chronicled the hardships faced by the migrant farm workers he saw as a child.

  Q: **In what town did Bill Aiken grow up?**

  A: **Madera** [But Google's BERT says <No Answer>!]

Why Choose The Default Final Project?

- If you:
  - Have limited experience with research, don't have any clear idea of what you want to do, or want guidance and a goal, … and a leaderboard, even
- Then:
  - Do the default final project! Many people should do it!

- Considerations:
  - The default final project gives you lots of guidance, scaffolding, and clear goalposts to aim at
  - The path to success is not to do something that looks kinda lame compared to what you could have done with the DFP

Why Choose The Custom Final Project?

- If you:
  - Have some research project that you're excited about (and are possibly already working on)
  - You want to try to do something different on your own
  - You're just interested in something other than question answering (that involves human language material and deep learning)
  - You want to see more of the process of defining a research goal, finding data and tools, and working out something you could do that is interesting, and how to evaluate it
- Then:
  - Do the custom final project!

Project Proposal – from everyone 5%

1. Find a relevant research paper for your topic
   - For DFP, a paper on the SQuAD leaderboard will do, but you might look elsewhere for interesting QA/reading comprehension work
2. Write a summary of that research paper and describe how you hope to use or adapt ideas from it
3. Write what you plan to work on and how you can innovate in your final project work
   - Suggest a good milestone to have achieved as a halfway point
4. Describe as needed, especially for Custom projects:
   - A project plan, relevant existing literature, the kind(s) of models you will use/explore; the data you will use (and how it is obtained), and how you will evaluate success

3–4 pages. Details released this Thursday

Due Tue Feb 11, 4:30pm on Gradescope

Project Proposal – from everyone 5%

1. How to think critically about a research paper
   - Grading of research paper review is primarily evaluative
   - What were the novel contributions or points?
   - Is what makes it work something general and reusable?
   - Are there flaws or neat details in what they did?
   - How does it fit with other papers on similar topics?
   - Does it provoke good questions on further or different things to try?
2. How to do a good job on your project proposal
   - Grading of project proposal is primarily formative
   - You need to have an overall sensible idea (!)
   - But most project plans that are lacking are lacking in nuts-and-bolts ways:
     - Do you have good data or a realistic plant to be able to collect it
     - Do you have a realistic way to evaluate your work
     - Do you have appropriate baselines or proposed ablation studies for comparisons

Project Milestone – from everyone 5%

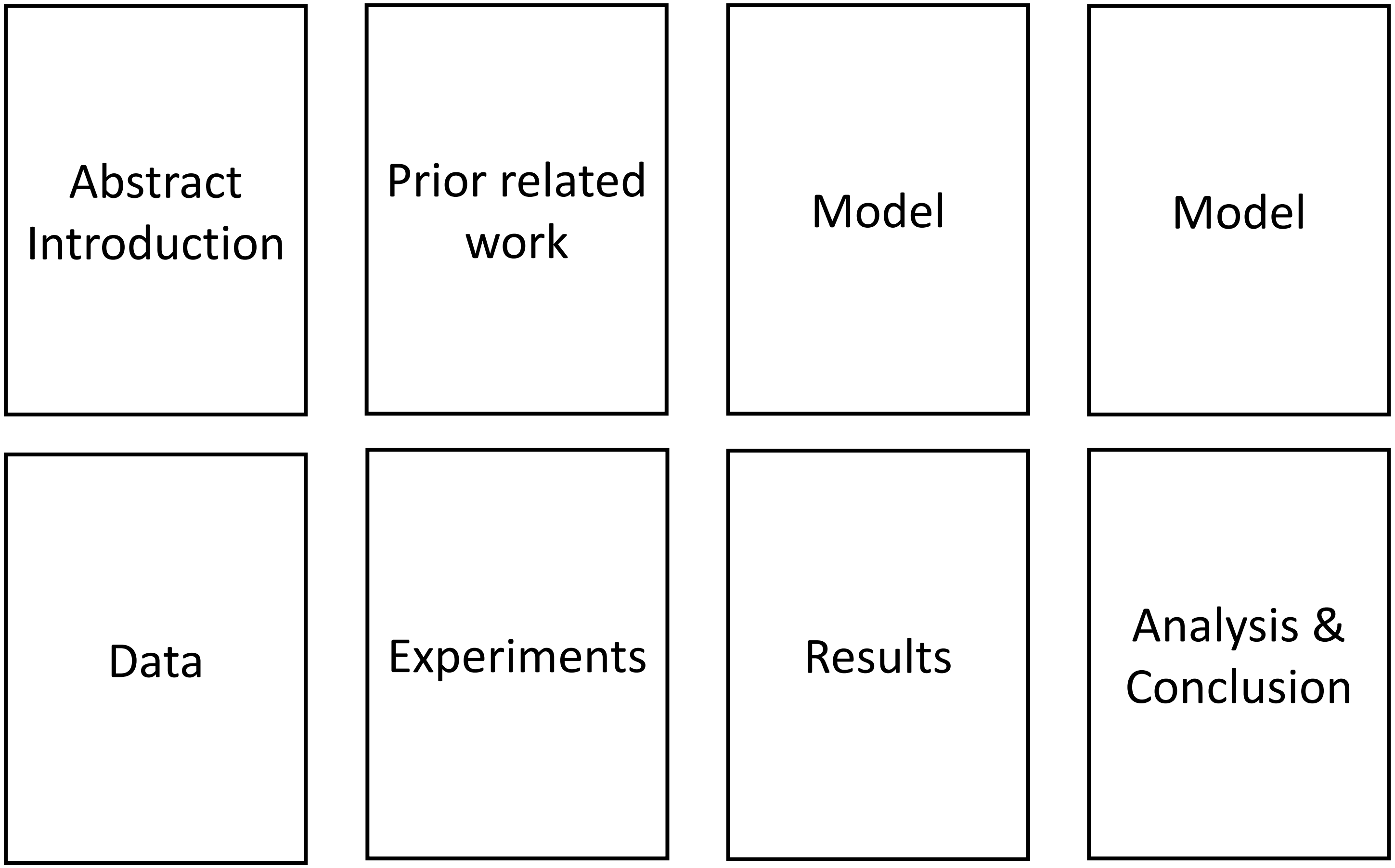
- This is a progress report
- You should be more than halfway done!
- Describe the experiments you have run
- Describe the preliminary results you have obtained
- Describe how you plan to spend the rest of your time

You are expected to **have implemented some system** and to **have some initial experimental results** to show by this date (except for certain unusual kinds of projects)

Due Tue Mar 3, 4:30pm on Gradescope

Project writeup

- Writeup quality is important to your grade!
  - Look at last-year's prize winners for examples



Abstract Introduction

Prior related work

Model

Model

Data

Experiments

Results

Analysis & Conclusion

Much of today's info is relevant ... for everybody

- At a lofty level
  - It's good to know something about how to do research!
- At a prosaic level
  - We'll touch on:
    - Baselines
    - Benchmarks
    - Evaluation
    - Error analysis
    - Paper writing

    which are all great things to know about for the DFP too!

2. Finding Research Topics

Two basic starting points, for all of science:

- [Nails] Start with a (domain) problem of interest and try to find good/better ways to address it than are currently known/used
- [Hammers] Start with a technical approach of interest, and work out good ways to extend or improve it or new ways to apply it

Project types

This is not an exhaustive list, but most projects are one of

1. Find an application/task of interest and explore how to approach/solve it effectively, often with an existing model
   - Could be task in the wild or some existing Kaggle/bake-off/shared task
2. Implement a complex neural architecture and demonstrate its performance on some data
3. Come up with a new or variant neural network model and explore its empirical success
4. Analysis project. Analyze the behavior of a model: how it represents linguistic knowledge or what kinds of phenomena it can handle or errors that it makes
5. Rare theoretical project: Show some interesting, non-trivial properties of a model type, data, or a data representation

Deep Poetry: Word-Level and Character-Level Language Models for Shakespearean Sonnet Generation

Stanley Xie, Ruchir Rastogi and Max Chang

Gated LSTM

Thy youth 's time and face his form shall cover?
Now all fresh beauty, my love there
Will ever Time to greet, forget each, like ever decease,
But in a best at worship his glory die.

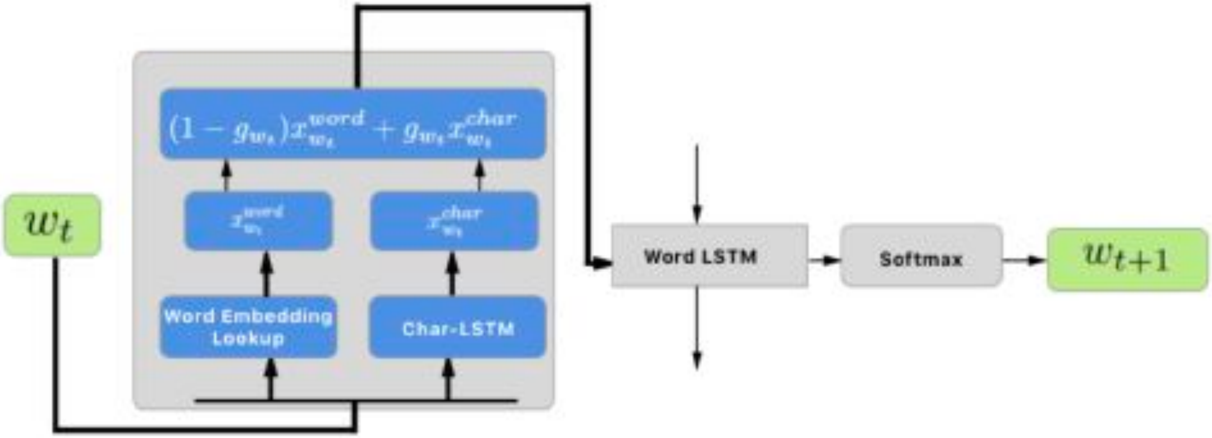
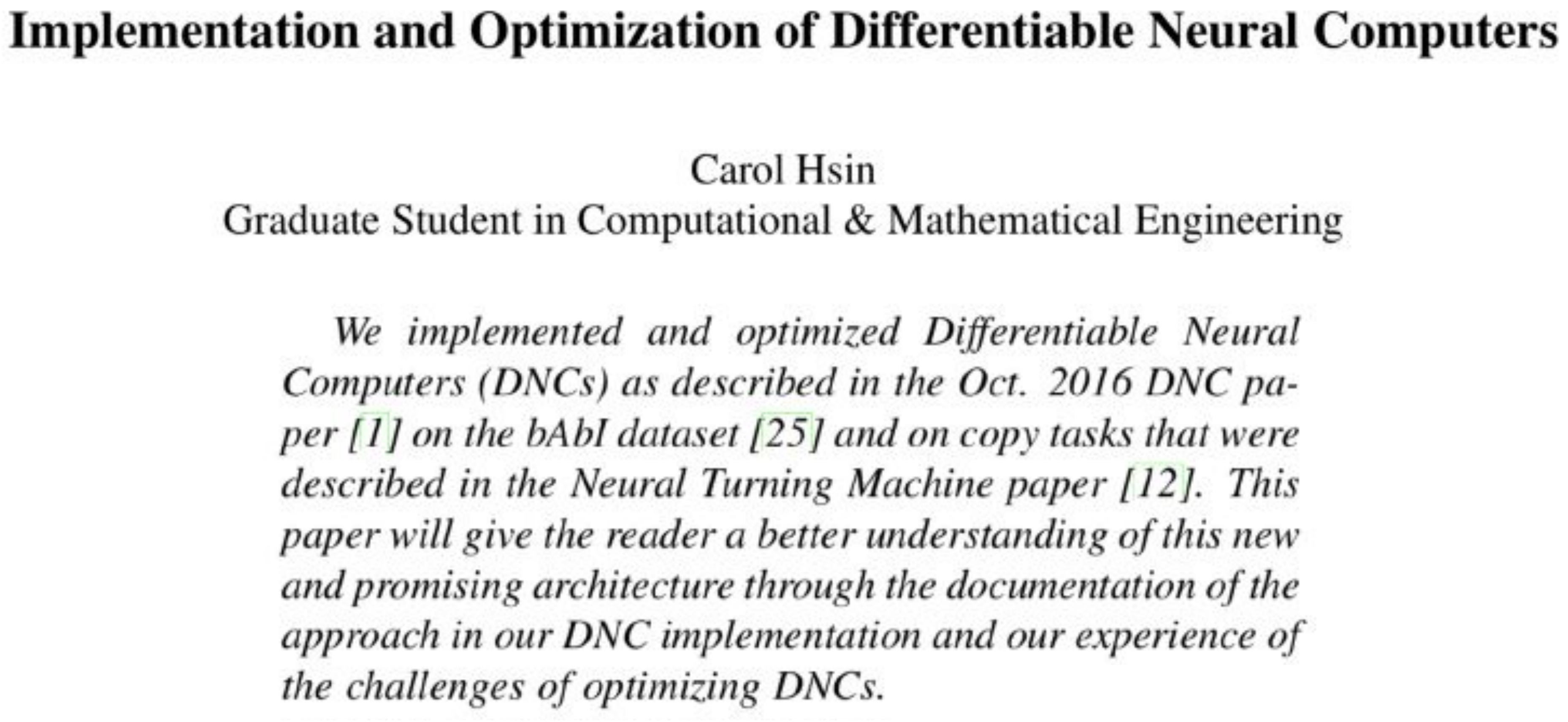


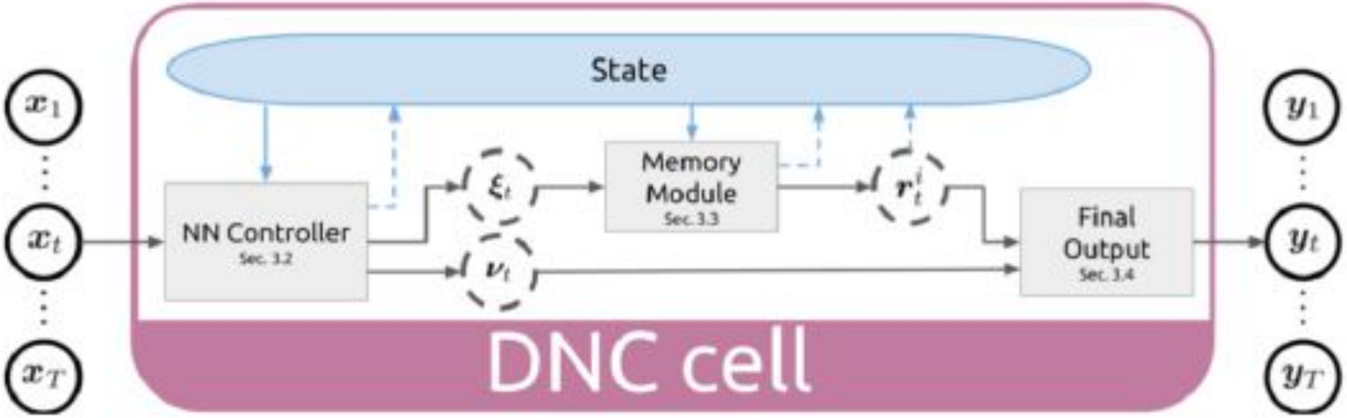
Figure 1: Architecture of the Gated LSTM

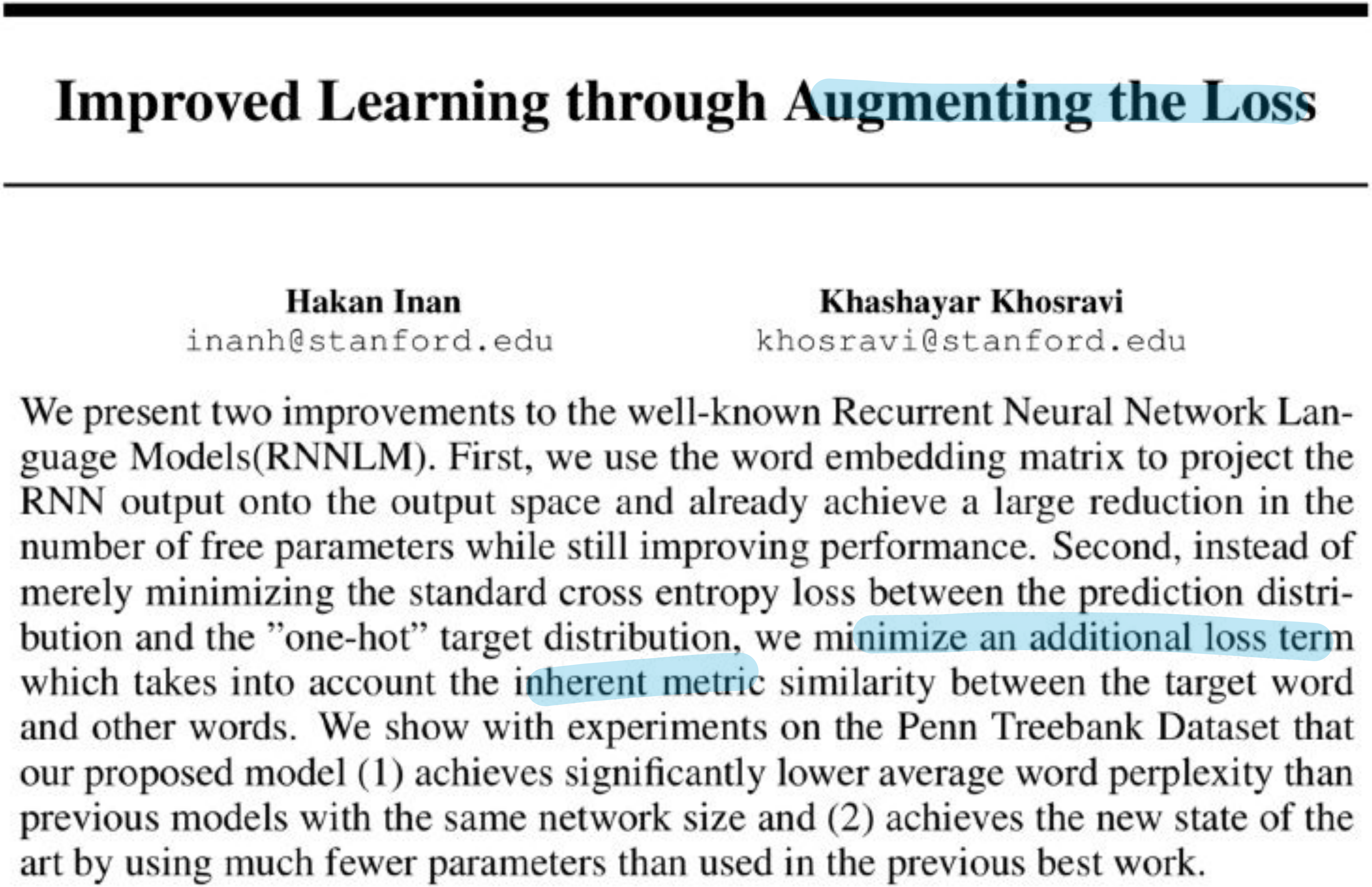


Implementation and Optimization of Differentiable Neural Computers

Carol Hsin
Graduate Student in Computational & Mathematical Engineering

We implemented and optimized Differentiable Neural Computers (DNCs) as described in the Oct. 2016 DNC paper [1] on the bAbI dataset [25] and on copy tasks that were described in the Neural Turning Machine paper [12]. This paper will give the reader a better understanding of this new and promising architecture through the documentation of the approach in our DNC implementation and our experience of the challenges of optimizing DNCs.





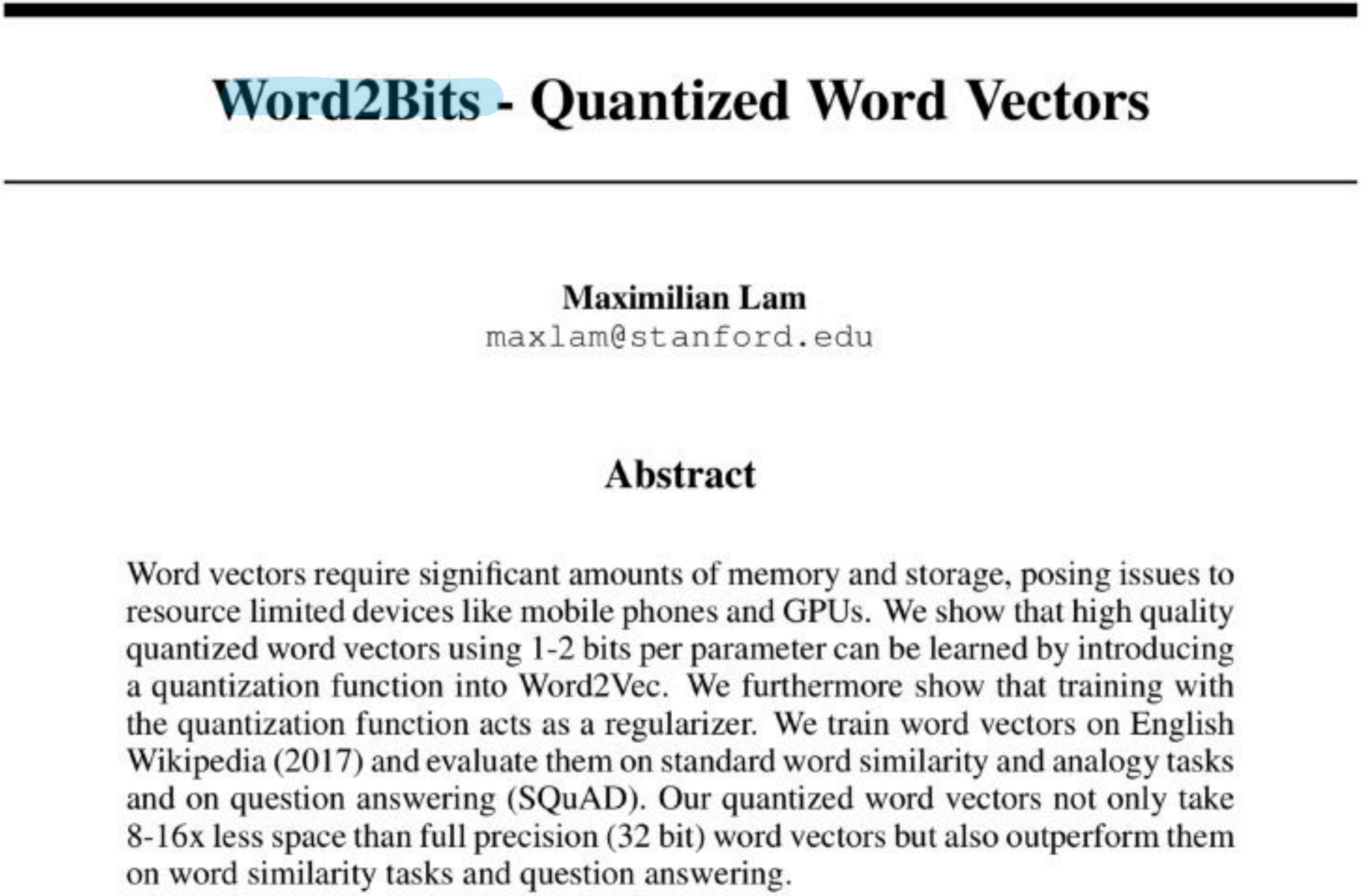
Improved Learning through Augmenting the Loss

Hakan Inan
inanh@stanford.edu

Khashayar Khosravi
khosravi@stanford.edu

We present two improvements to the well-known Recurrent Neural Network Language Models(RNNLM). First, we use the word embedding matrix to project the RNN output onto the output space and already achieve a large reduction in the number of free parameters while still improving performance. Second, instead of merely minimizing the standard cross entropy loss between the prediction distribution and the ”one-hot” target distribution, we minimize an additional loss term which takes into account the inherent metric similarity between the target word and other words. We show with experiments on the Penn Treebank Dataset that our proposed model (1) achieves significantly lower average word perplexity than previous models with the same network size and (2) achieves the new state of the art by using much fewer parameters than used in the previous best work.

Published as a conference paper at ICLR 2017



Word2Bits - Quantized Word Vectors

Maximilian Lam
maxlam@stanford.edu

Abstract

Word vectors require significant amounts of memory and storage, posing issues to resource limited devices like mobile phones and GPUs. We show that high quality quantized word vectors using 1-2 bits per parameter can be learned by introducing a quantization function into Word2Vec. We furthermore show that training with the quantization function acts as a regularizer. We train word vectors on English Wikipedia (2017) and evaluate them on standard word similarity and analogy tasks and on question answering (SQuAD). Our quantized word vectors not only take 8-16x less space than full precision (32 bit) word vectors but also outperform them on word similarity tasks and question answering.

How to find an interesting place to start?

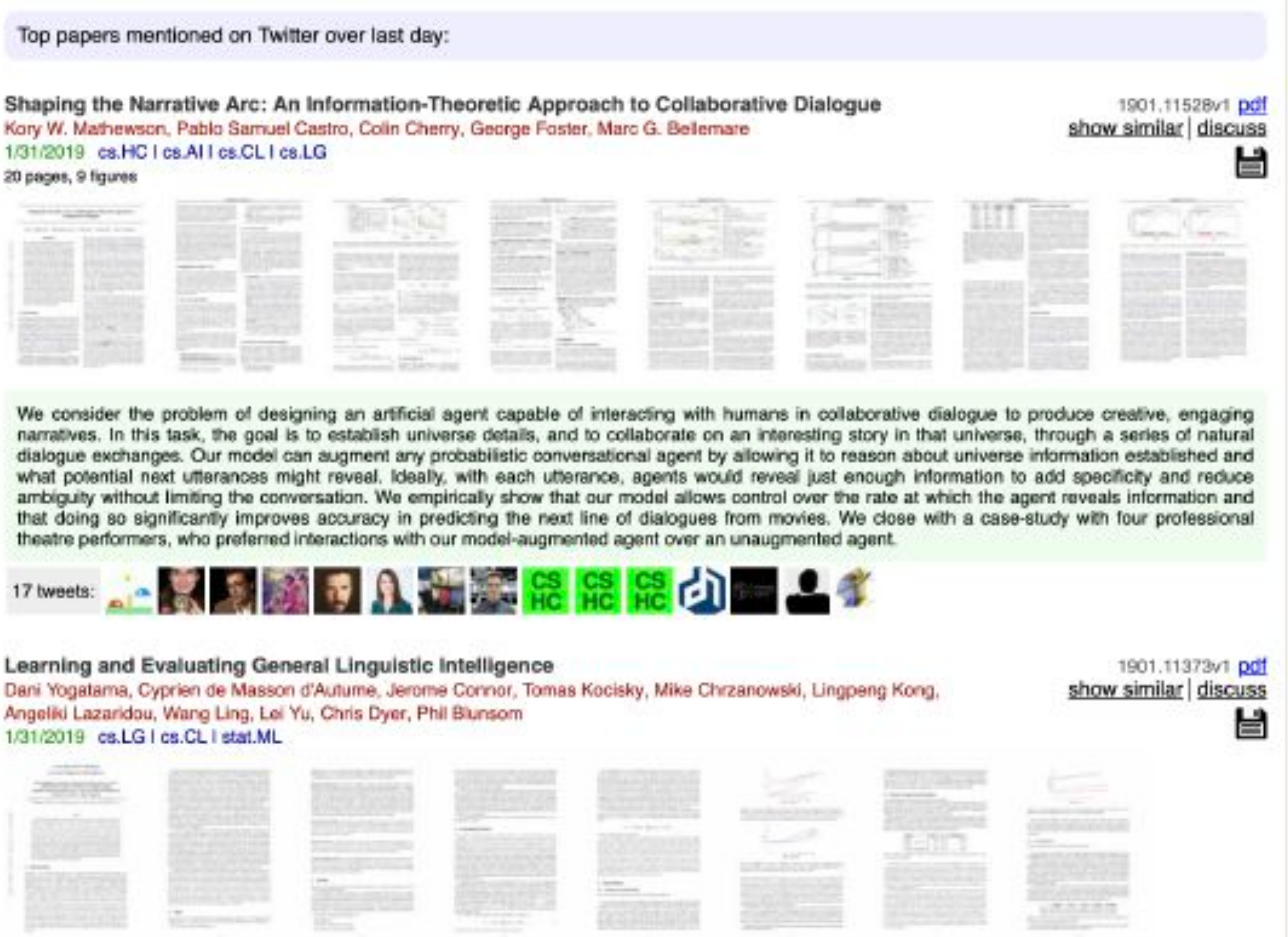
- Look at ACL anthology for NLP papers:
  - https://aclanthology.info
- Also look at the online proceedings of major ML conferences:
  - NeurIPS, ICML, ICLR
- Look at past cs224n projects
  - See the class website
- Look at online preprint servers, especially:
  - https://arxiv.org

- Even better: look for an interesting problem in the world

How to find an interesting place to start?

Arxiv Sanity Preserver by Stanford grad Andrej Karpathy of cs231n

http://www.arxiv-sanity.com



Top papers mentioned on Twitter over last day:

Shaping the Narrative Arc: An Information-Theoretic Approach to Collaborative Dialogue

Kory W. Mathewson, Pablo Samuel Castro, Colin Cherry, George Foster, Marc G. Bellemare

1/31/2019 cs.HC | cs.AI | cs.CL | cs.LG

20 pages, 9 figures

1901.11528v1 pdf

show similar | discuss

We consider the problem of designing an artificial agent capable of interacting with humans in collaborative dialogue to produce creative, engaging narratives. In this task, the goal is to establish universe details, and to collaborate on an interesting story in that universe, through a series of natural dialogue exchanges. Our model can augment any probabilistic conversational agent by allowing it to reason about universe information established and what potential next utterances might reveal. Ideally, with each utterance, agents would reveal just enough information to add specificity and reduce ambiguity without limiting the conversation. We empirically show that our model allows control over the rate at which the agent reveals information and that doing so significantly improves accuracy in predicting the next line of dialogues from movies. We close with a case-study with four professional theatre performers, who preferred interactions with our model-augmented agent over an unaugmented agent.

17 tweets:

Learning and Evaluating General Linguistic Intelligence

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, Phil Blunsom

1/31/2019 cs.LG | cs.CL | stat.ML

1901.11373v1 pdf

show similar | discuss

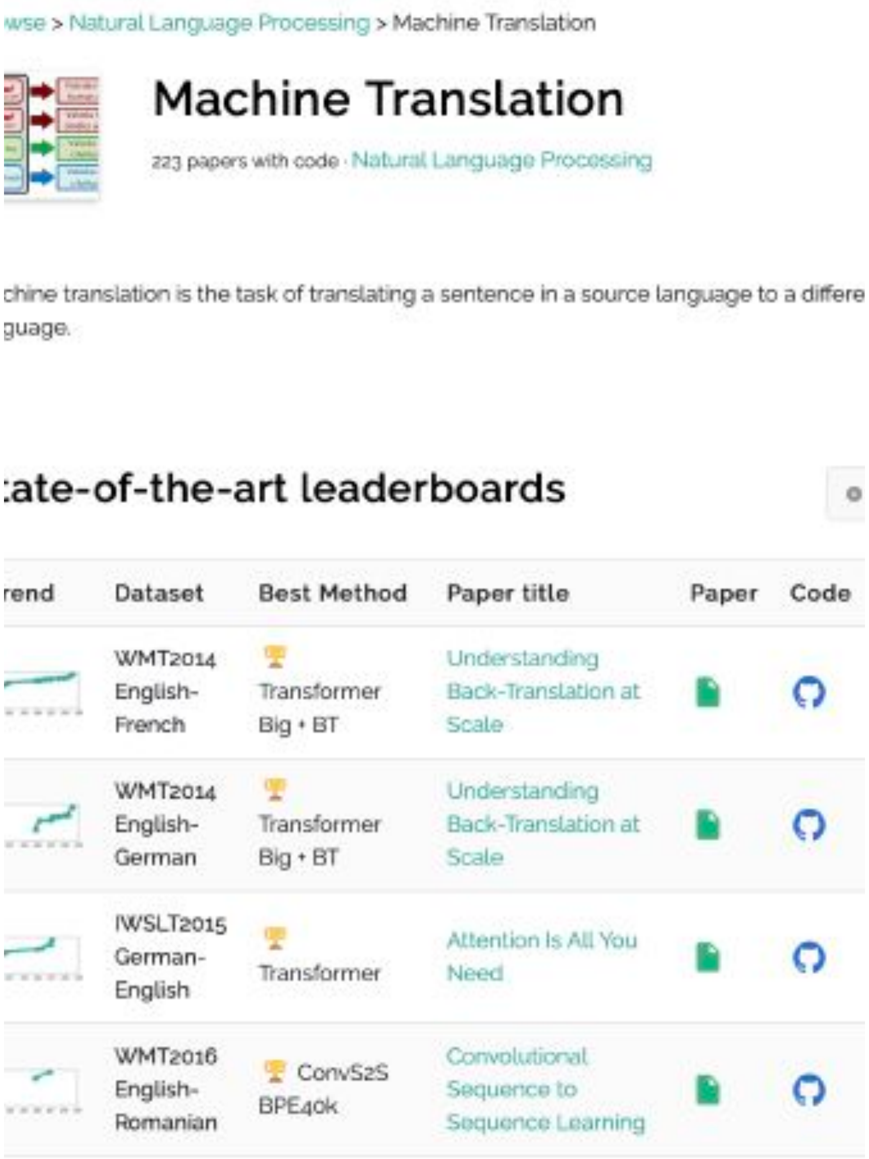
Want to beat the state of the art on something?

Great new sites that try to collate info on the state of the art

- Not always correct, though

https://paperswithcode.com/sota
https://nlpprogress.com/
https://github.com/RedditSota/state-of-the-art-result-for-machine-learning-problems/

https://gluebenchmark.com/leaderboard/
https://www.conll.org/previous-tasks/



wse > Natural Language Processing > Machine Translation

Machine Translation

223 papers with code · Natural Language Processing

chine translation is the task of translating a sentence in a source language to a differe
guage.

tate-of-the-art leaderboards

rend	Dataset	Best Method	Paper title	Paper	Code
	WMT2014 English-French	🏆 Transformer Big + BT	Understanding Back-Translation at Scale	📄	
	WMT2014 English-German	🏆 Transformer Big + BT	Understanding Back-Translation at Scale	📄	
	IWSLT2015 German-English	🏆 Transformer	Attention Is All You Need	📄	
	WMT2016 English-Romanian	🏆 ConvS2S BPE40k	Convolutional Sequence to Sequence Learning	📄	

Finding a topic

- Turing award winner and Stanford CS emeritus professor Ed Feigenbaum says to follow the advice of his advisor, AI pioneer, and Turing and Nobel prize winner Herb Simon:
  - “If you see a research area where many people are working, go somewhere else.”
- But where to go? Wayne Gretzky:
  - “I skate to where the puck is going, not where it has been.”

Must-haves for most* custom final projects

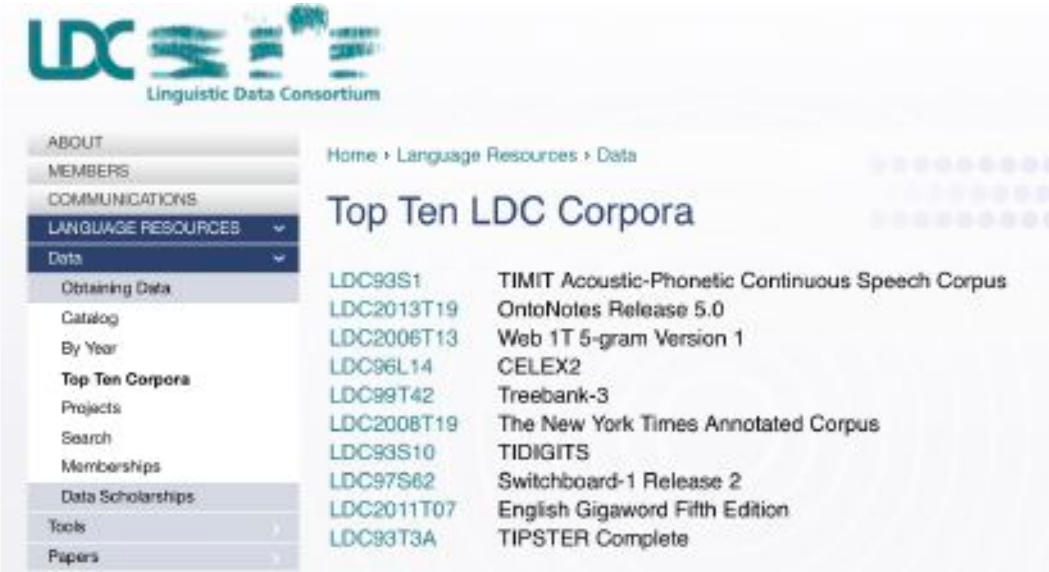
- Suitable data
  - Usually aiming at: 10,000+ labeled examples by milestone
- Feasible task
- Automatic evaluation metric
- Human language is central to the project
- You use some neural networks/deep learning

3. Finding data

- Some people collect their own data for a project – **we like that!**
  - You may have a project that uses "unsupervised" data
  - You can annotate a small amount of data
  - You can find a website that effectively provides annotations, such as likes, stars, ratings, responses, etc.
    - Let's you learn about real word challenges of applying ML/NLP!
- Some people have existing data from a research project or company
  - Fine to use providing you can provide data samples for submission, report, etc.
- **Most people make use of an existing, curated dataset built by previous researchers**
    - You get a fast start and there is obvious prior work and baselines

Linguistic Data Consortium

- https://catalog.ldc.upenn.edu/
- Stanford licenses data; you can get access by signing up at: https://linguistics.stanford.edu/resources/resources-corpora
- Treebanks, named entities, coreference data, lots of newswire, lots of speech with transcription, parallel MT data
  - Look at their catalog
  - Don't use for non-Stanford purposes!

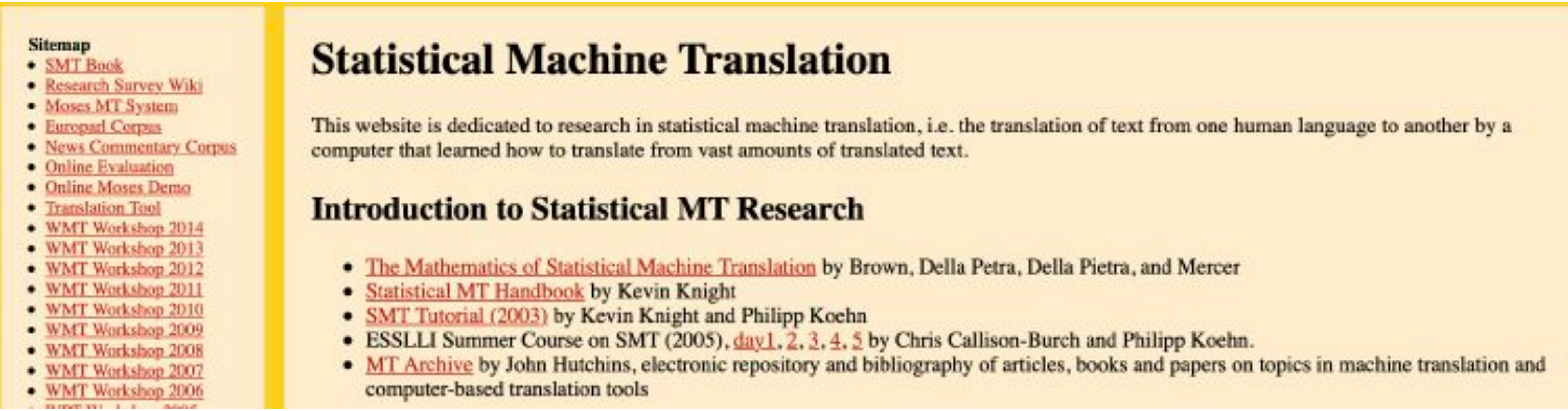


Top Ten LDC Corpora

LDC93S1	TIMIT Acoustic-Phonetic Continuous Speech Corpus
LDC2013T19	OntoNotes Release 5.0
LDC2006T13	Web 1T 5-gram Version 1
LDC96L14	CELEX2
LDC99T42	Treebank-3
LDC2008T19	The New York Times Annotated Corpus
LDC93S10	TIDIGITS
LDC97S62	Switchboard-1 Release 2
LDC2011T07	English Gigaword Fifth Edition
LDC93T3A	TIPSTER Complete

Machine translation

- http://statmt.org
- Look in particular at the various WMT shared tasks



Sitemap

- SMT Book
- Research Survey Wiki
- Moses MT System
- Europarl Corpus
- News Commentary Corpus
- Online Evaluation
- Online Moses Demo
- Translation Tool
- WMT Workshop 2014
- WMT Workshop 2013
- WMT Workshop 2012
- WMT Workshop 2011
- WMT Workshop 2010
- WMT Workshop 2009
- WMT Workshop 2008
- WMT Workshop 2007
- WMT Workshop 2006

Statistical Machine Translation

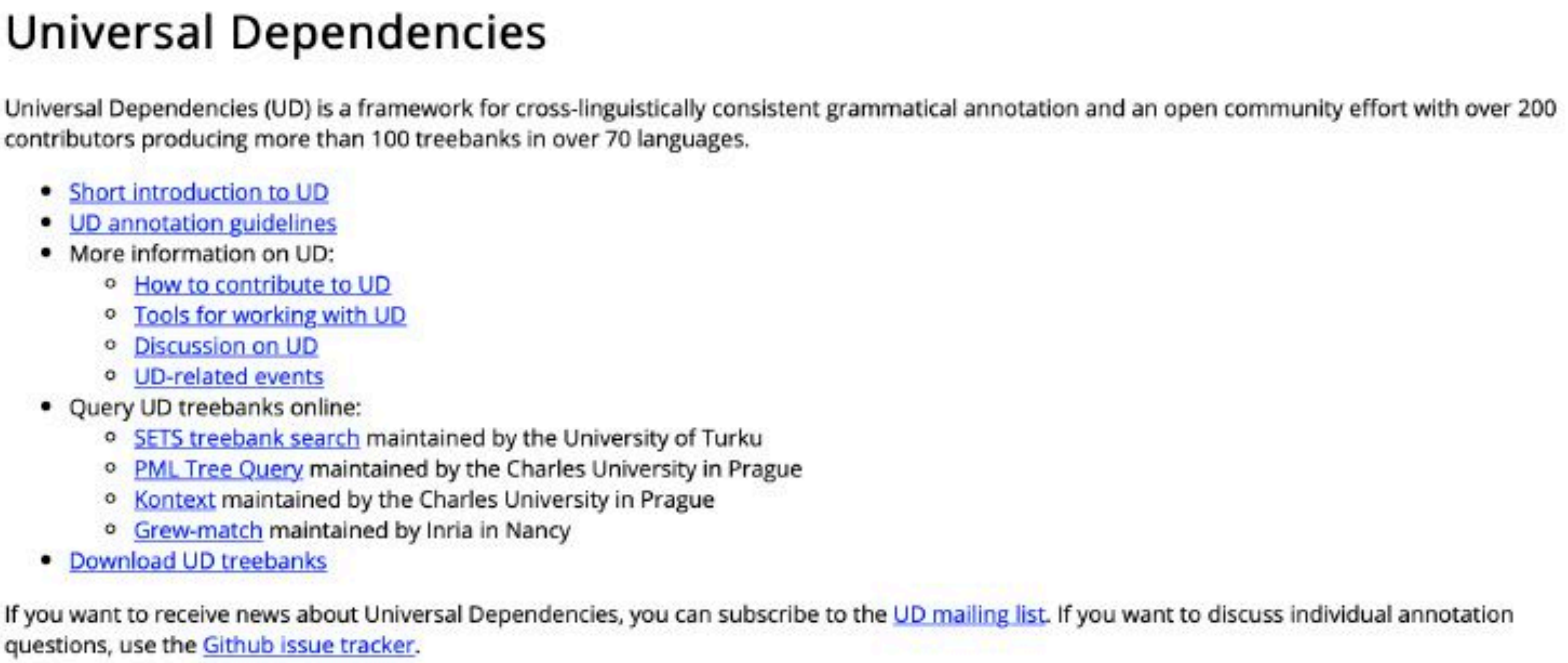
This website is dedicated to research in statistical machine translation, i.e. the translation of text from one human language to another by a computer that learned how to translate from vast amounts of translated text.

Introduction to Statistical MT Research

- The Mathematics of Statistical Machine Translation by Brown, Della Petra, Della Pietra, and Mercer
- Statistical MT Handbook by Kevin Knight
- SMT Tutorial (2003) by Kevin Knight and Philipp Koehn
- ESSLLI Summer Course on SMT (2005), day1, 2, 3, 4, 5 by Chris Callison-Burch and Philipp Koehn.
- MT Archive by John Hutchins, electronic repository and bibliography of articles, books and papers on topics in machine translation and computer-based translation tools

Dependency parsing: Universal Dependencies

- https://universaldependencies.org



Universal Dependencies

Universal Dependencies (UD) is a framework for cross-linguistically consistent grammatical annotation and an open community effort with over 200 contributors producing more than 100 treebanks in over 70 languages.

- Short introduction to UD
- UD annotation guidelines
- More information on UD:
  - How to contribute to UD
  - Tools for working with UD
  - Discussion on UD
  - UD-related events
- Query UD treebanks online:
  - SETS treebank search maintained by the University of Turku
  - PML Tree Query maintained by the Charles University in Prague
  - Kontext maintained by the Charles University in Prague
  - Grew-match maintained by Inria in Nancy
- Download UD treebanks

If you want to receive news about Universal Dependencies, you can subscribe to the UD mailing list. If you want to discuss individual annotation questions, use the Github issue tracker.

Many, many more

- There are now many other datasets available online for all sorts of purposes
  - Look at Kaggle
  - Look at research papers
  - Look at lists of datasets
    - https://machinelearningmastery.com/datasets-natural-language-processing/
    - https://github.com/niderhoff/nlp-datasets
  - Lots of particular things:
    - https://gluebenchmark.com/tasks
    - https://nlp.stanford.edu/sentiment/
    - https://research.fb.com/downloads/babi/ (Facebook bAbI-related)
  - Ask on Piazza or talk to course staff. Look at papers!

4. Doing your research example:
Straightforward Class Project: Apply NNets to Task

1. Define Task:
   - Example: **Summarization**

2. Define Dataset
   1. Search for academic datasets
      - They already have baselines
      - E.g.: Newsroom Summarization Dataset: https://summari.es

   2. Define your own data (harder, need new baselines)
      - Allows connection to your research
      - A fresh problem provides fresh opportunities!
      - Be creative: Twitter, Blogs, News, etc. There are lots of neat websites which provide creative opportunities for new tasks

Straightforward Class Project: Apply NNets to Task

3. Dataset hygiene
   - Right at the beginning, separate off devtest and test splits
     - Discussed more next

4. Define your metric(s)
   - Search online for well established metrics on this task
   - Summarization: Rouge (Recall-Oriented Understudy for Gisting Evaluation) which defines *n*-gram overlap to human summaries
   - Human evaluation is still much better for summarization; you may be able to do a small scale human eval

Straightforward Class Project: Apply NNets to Task

5. Establish a baseline
   - Implement the simplest model first (often logistic regression on unigrams and bigrams or averaging word vectors)
     - For summarization: See LEAD-3 baseline
   - Compute metrics on train AND dev NOT test
   - Analyze errors
   - If metrics are amazing and no errors:
     - Done! Problem was too easy. Need to restart. ☺/☹

6. Implement existing neural net model
   - Compute metric on train and dev
   - Analyze output and errors
   - Minimum bar for this class

Straightforward Class Project: Apply NNets to Task

7. Always be close to your data! (Except for the final test set!)
    - Visualize the dataset
    - Collect summary statistics
    - Look at errors
    - Analyze how different hyperparameters affect performance

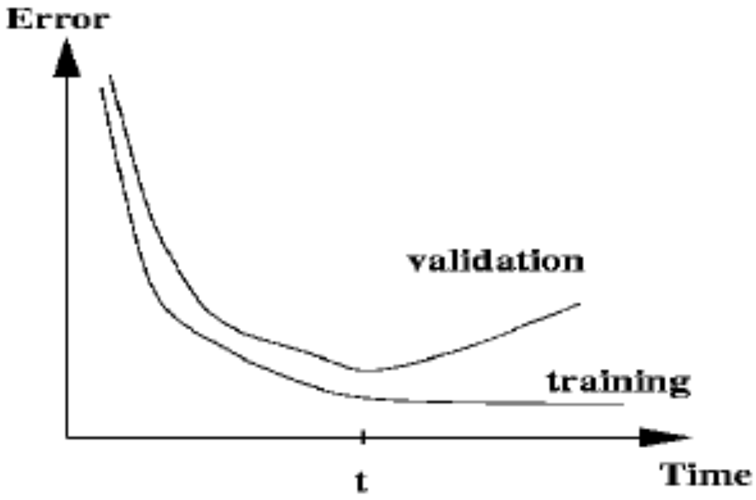
8. Try out different models and model variants
   Aim to iterate quickly via having a good experimental setup
    - Fixed window neural model
    - Recurrent neural network
    - Recursive neural network
    - Convolutional neural network
    - Attention-based model
    - …

Pots of data

- Many publicly available datasets are released with a **train/dev/test** structure. **We're all on the honor system to do test-set runs only when development is complete.**
- Splits like this presuppose a fairly large dataset.
- If there is no dev set or you want a separate tune set, then you create one by splitting the training data, though you have to weigh its size/usefulness against the reduction in train-set size.
- Having a fixed test set ensures that all systems are assessed against the same gold data. This is generally good, but it is problematic where the test set turns out to have unusual properties that distort progress on the task.

Training models and pots of data

- When training, models **overfit** to what you are training on
  - The model correctly describes what happened to occur in particular data you trained on, but the patterns are not general enough patterns to be likely to apply to new data
- The way to monitor and avoid problematic overfitting is using **independent** validation and test sets …



Training models and pots of data

- You build (estimate/train) a model on a **training set**.
- Often, you then set further hyperparameters on another, independent set of data, the **tuning set**
  - The tuning set is the training set for the hyperparameters!
- You measure progress as you go on a **dev set** (development test set or validation set)
  - If you do that a lot you overfit to the dev set so it can be good to have a second dev set, the **dev2** set
- **Only at the end**, you evaluate and present final numbers on a **test set**
  - Use the final test set **extremely** few times … ideally only once

Training models and pots of data

- The **train**, **tune**, **dev**, and **test** sets need to be completely distinct
- It is invalid to test on material you have trained on
  - You will get a falsely good performance. We usually overfit on train
- You need an independent tuning set
  - The hyperparameters won't be set right if tune is same as train
- If you keep running on the same evaluation set, you begin to overfit to that evaluation set
  - Effectively you are "training" on the evaluation set … you are learning things that do and don't work on that particular eval set and using the info
- To get a valid measure of system performance you need another untrained on, **independent** test set … hence dev2 and final test

Getting your neural network to train

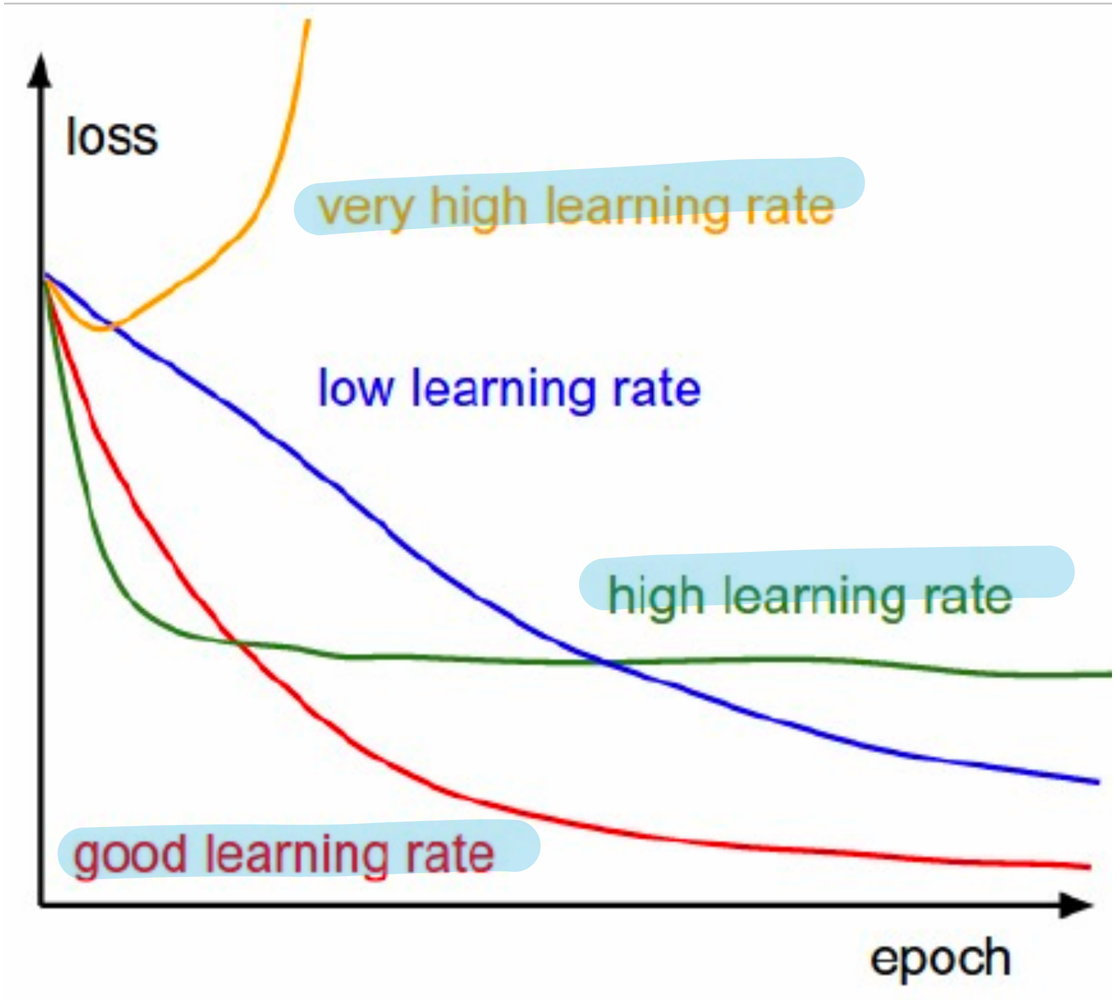
- Start with a positive attitude!
  - **Neural networks want to learn!**
    - If the network isn’t learning, you’re doing something to prevent it from learning successfully

- Realize the grim reality:
  - **There are lots of things that can cause neural nets to not learn at all or to not learn very well**
    - Finding and fixing them (“debugging and tuning”) can often take more time than implementing your model

- It’s hard to work out what these things are
  - But experience, experimental care, and rules of thumb help!

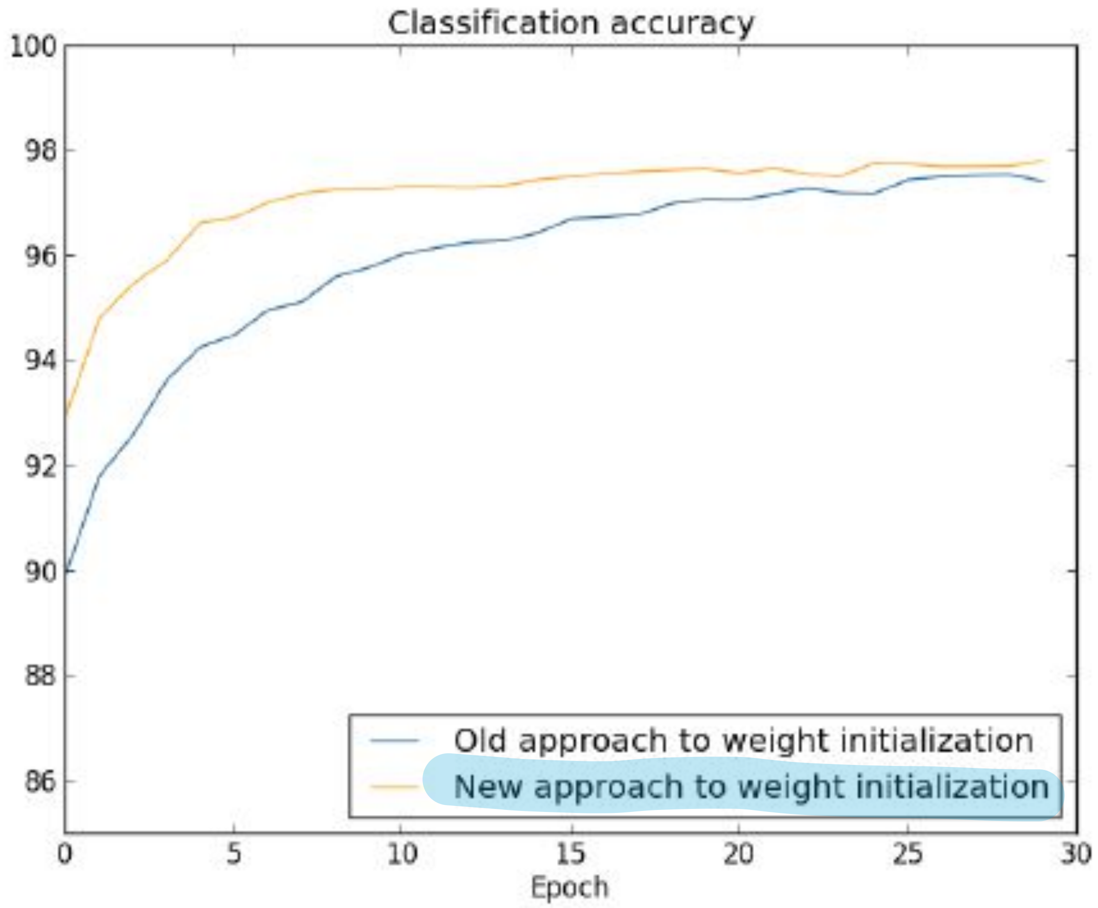
Models are sensitive to learning rates

- From Andrej Karpathy, CS231n course notes



Models are sensitive to initialization

- From Michael Nielsen http://neuralnetworksanddeeplearning.com/chap3.html



Training a (gated) RNN

1. Use an LSTM or GRU: *it makes your life so much simpler!*
2. Initialize recurrent matrices to be orthogonal
3. Initialize other matrices with a sensible (**small!**) scale
4. Initialize forget gate bias to 1: *default to remembering*
5. Use adaptive learning rate algorithms: *Adam, AdaDelta, …*
6. Clip the norm of the gradient: *1–5 seems to be a reasonable threshold when used together with Adam or AdaDelta.*
7. Either only dropout vertically or look into using Bayesian Dropout (Gal and Gahramani – not natively in PyTorch)
8. *Be patient! Optimization takes time*

[Saxe et al., ICLR2014;
Ba, Kingma, ICLR2015;
Zeiler, arXiv2012;
Pascanu et al., ICML2013]

Experimental strategy

- Work incrementally!
- Start with a very simple model and get it to work!
  - It’s hard to fix a complex but broken model
- Add bells and whistles one-by-one and get the model working with each of them (or abandon them)

- Initially run on a tiny amount of data
  - You will see bugs much more easily on a tiny dataset
  - Something like 4–8 examples is good
  - Often synthetic data is useful for this
  - Make sure you can get 100% on this data
    - Otherwise your model is definitely either not powerful enough or it is broken

Experimental strategy

- Run your model on a large dataset
  - It should still score close to 100% on the training data after optimization
    - Otherwise, you probably want to consider a more powerful model
    - Overfitting to training data is **not** something to be scared of when doing deep learning
      - These models are usually good at generalizing because of the way distributed representations share statistical strength regardless of overfitting to training data
- But, still, you now want good generalization performance:
  - Regularize your model until it doesn't overfit on dev data
    - Strategies like L2 regularization can be useful
    - But normally **generous dropout** is the secret to success

Details matter!

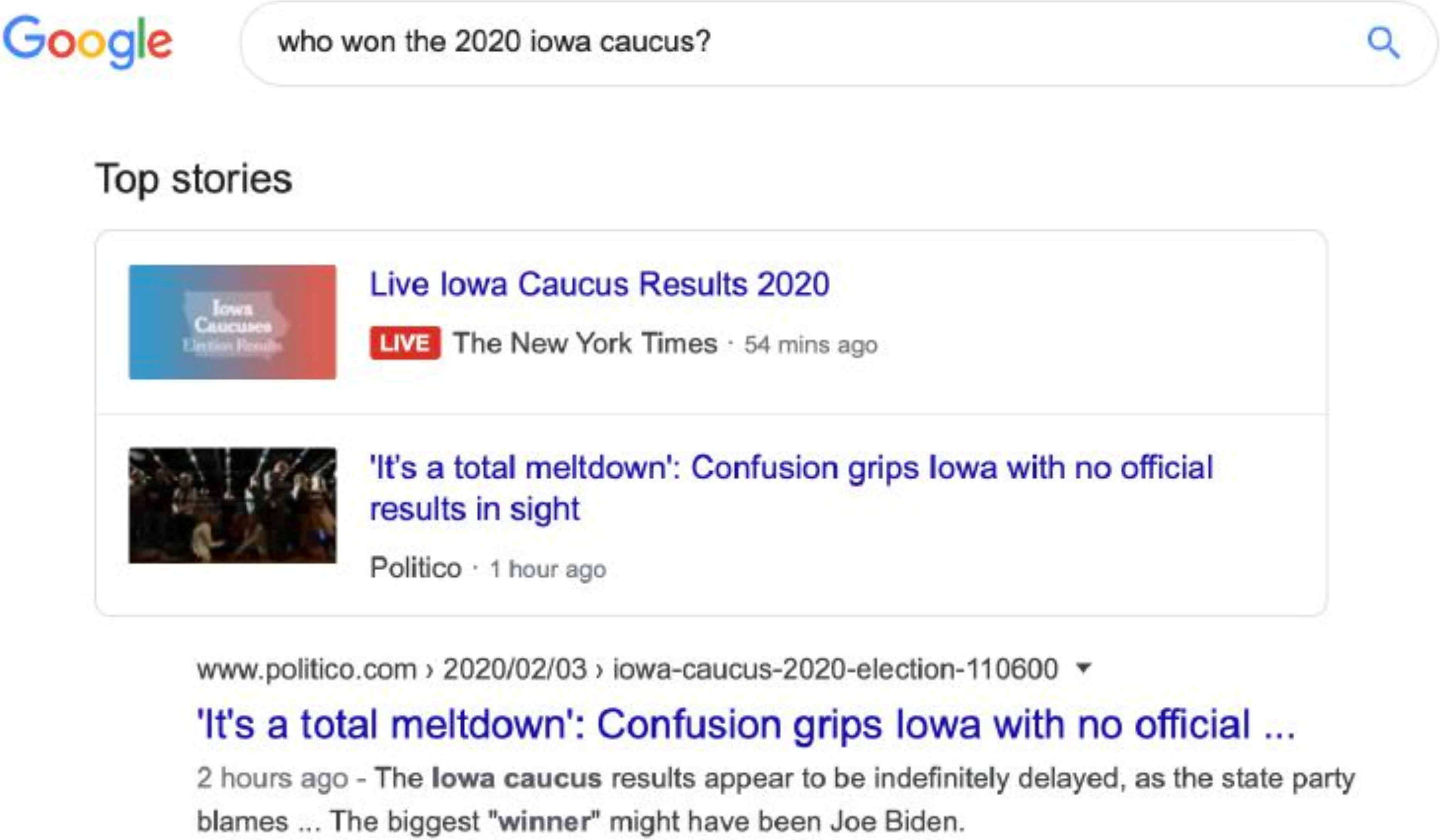
- Look at your data, collect summary statistics
- Look at your model's outputs, do error analysis
- Tuning hyperparameters is **really** important to almost all of the successes of NNets

The Default Final Project

Reading Comprehension

a.k.a. Question Answering

over documents



who won the 2020 iowa caucus?

Top stories

Live Iowa Caucus Results 2020

LIVE The New York Times · 54 mins ago

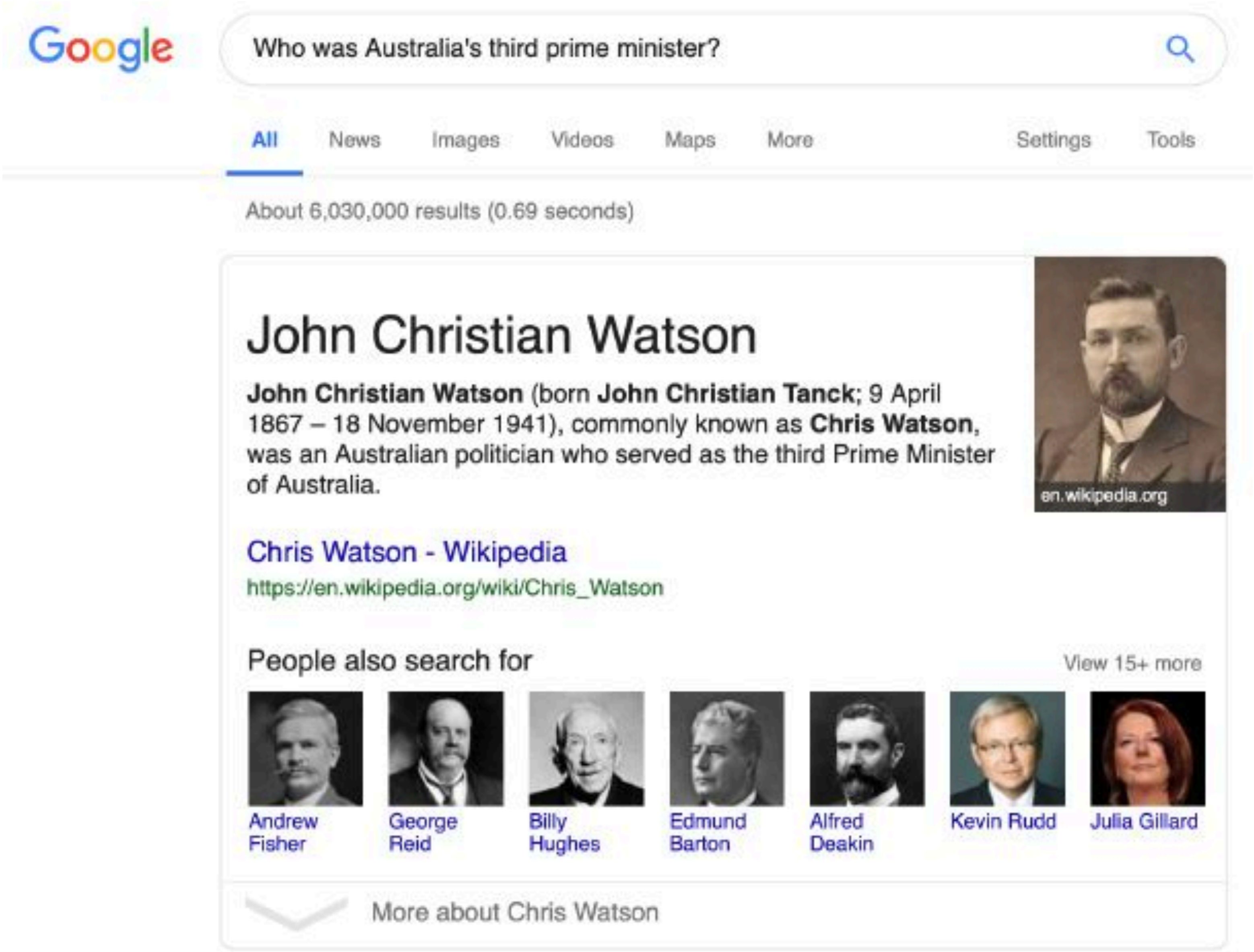
'It's a total meltdown': Confusion grips Iowa with no official results in sight

Politico · 1 hour ago

www.politico.com › 2020/02/03 › iowa-caucus-2020-election-110600

'It's a total meltdown': Confusion grips Iowa with no official ...

2 hours ago - The **Iowa caucus** results appear to be indefinitely delayed, as the state party blames ... The biggest "**winner**" might have been Joe Biden.



Who was Australia's third prime minister?

About 6,030,000 results (0.69 seconds)

John Christian Watson

John Christian Watson (born **John Christian Tanck**; 9 April 1867 – 18 November 1941), commonly known as **Chris Watson**, was an Australian politician who served as the third Prime Minister of Australia.

en.wikipedia.org

Chris Watson - Wikipedia

https://en.wikipedia.org/wiki/Chris_Watson

People also search for

View 15+ more

Andrew Fisher, George Reid, Billy Hughes, Edmund Barton, Alfred Deakin, Kevin Rudd, Julia Gillard

More about Chris Watson

Technical note: This is a “featured snippet” answer extracted from a web page, not a question answered using the (structured) Google Knowledge Graph (formerly known as Freebase).

2. Motivation: Question answering

- With massive collections of full-text documents, i.e., the web ☺, simply returning relevant documents is of limited use
- Rather, we often want **answers** to our **questions**
- Especially on mobile
- Or using a digital assistant device, like Alexa, Google Assistant, …

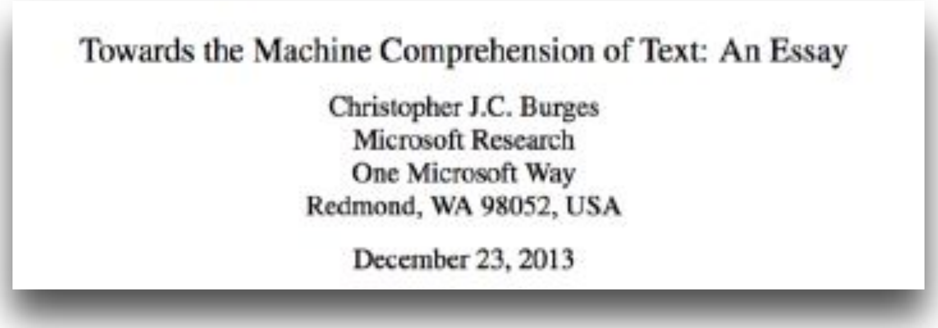
- We can factor this into two parts:
  1. Finding documents that (might) contain an answer
     - Which can be handled by traditional information retrieval/web search
     - (I teach cs276 which deals with this problem)
  2. Finding an answer in a paragraph or a document
     - This problem is often termed **Reading Comprehension**
     - It is what we will focus on today

A Brief History of Reading Comprehension

- Much early NLP work attempted reading comprehension
  - Schank, Abelson, Lehnert et al. c. 1977 – “Yale A.I. Project”
- Revived by Lynette Hirschman in 1999:
  - Could NLP systems answer human reading comprehension questions for 3rd to 6th graders? Simple methods attempted.
- Revived again by Chris Burges in 2013 with MCTest
  - Again answering questions over simple story texts
- Floodgates opened in 2015/16 with the production of large datasets which permit supervised neural systems to be built
  - Hermann et al. (NIPS 2015) DeepMind CNN/DM dataset
  - Rajpurkar et al. (EMNLP 2016) SQuAD
  - MS MARCO, TriviaQA, RACE, NewsQA, NarrativeQA, …

Machine Comprehension (Burges 2013)

- "A machine **comprehends** a passage of **text** if, for any **question** regarding that text that can be **answered** correctly by a majority of native speakers, that machine can provide a string which those speakers would agree both answers that question, and does not contain information irrelevant to that question."



Towards the Machine Comprehension of Text: An Essay

Christopher J.C. Burges
Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA

December 23, 2013

MCTest Reading Comprehension

Passage (*P*) + Question (*Q*) → Answer (*A*)

P

Alyssa got to the beach after a long trip. She's from Charlotte. She traveled from Atlanta. She's now in Miami. She went to Miami to visit some friends. But she wanted some time to herself at the beach, so she went there first. After going swimming and laying out, she went to her friend Ellen's house. Ellen greeted Alyssa and they both had some lemonade to drink. Alyssa called her friends Kristin and Rachel to meet at Ellen's house…….

Q Why did Alyssa go to Miami?

A To visit some friends

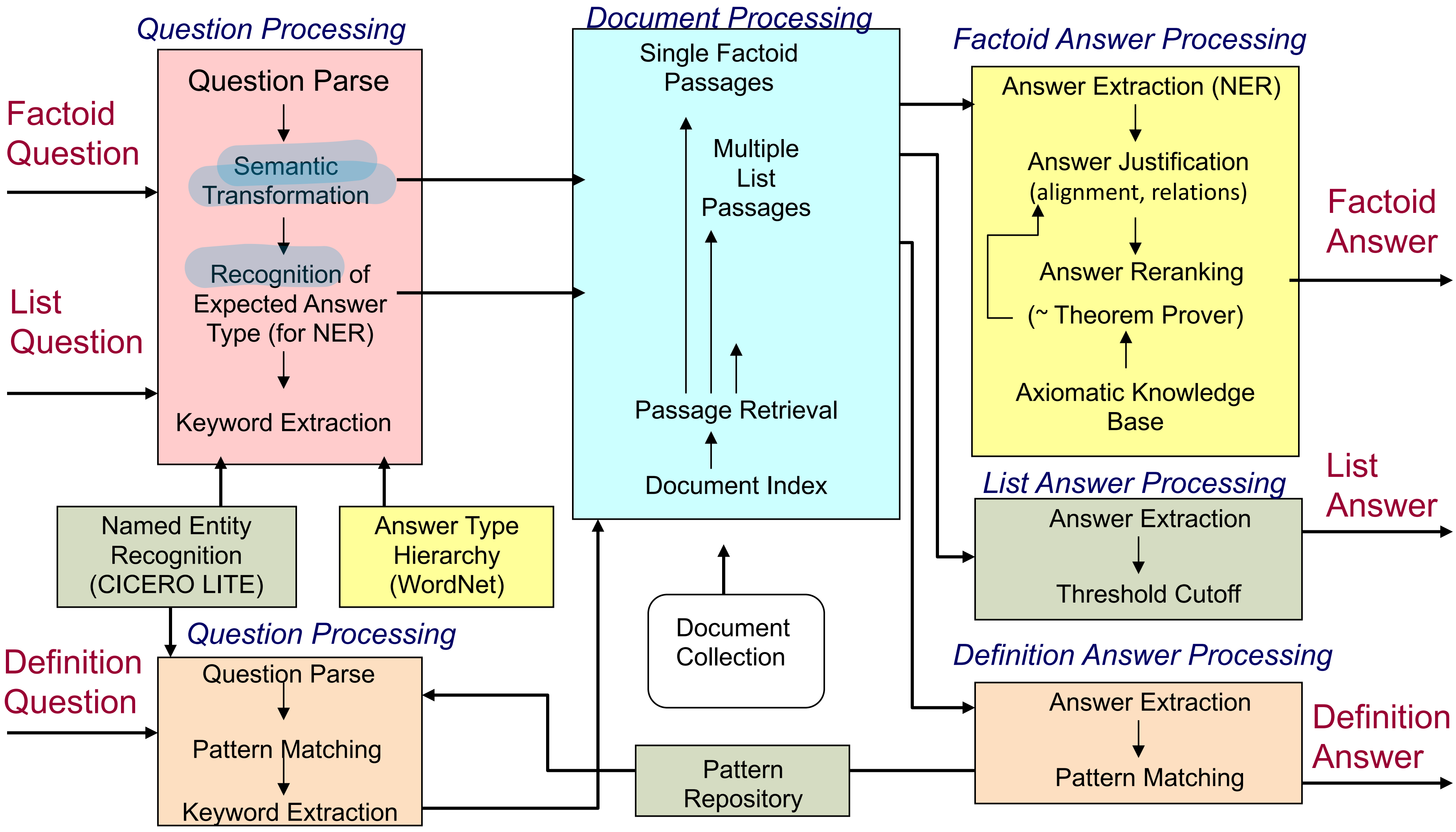
A Brief History of Open-domain Question Answering

- Simmons et al. (1964) did first exploration of answering questions from an expository text based on matching dependency parses of a question and answer
- Murax (Kupiec 1993) aimed to answer questions over an online encyclopedia using IR and shallow linguistic processing
- The NIST TREC QA track begun in 1999 first rigorously investigated answering fact questions over a large collection of documents
- IBM’s Jeopardy! System (DeepQA, 2011) brought attention to a version of the problem; it used an ensemble of many methods
- DrQA (Chen et al. 2016) uses IR followed by neural reading comprehension to bring deep learning to Open-domain QA

Turn-of-the Millennium Full NLP QA:

[architecture of LCC (Harabagiu/Moldovan) QA system, circa 2003]

Complex systems but they did work fairly well on “factoid” questions



3. Stanford Question Answering Dataset (SQuAD)

(Rajpurkar et al., 2016)

Question: Which team won Super Bowl 50?

Passage

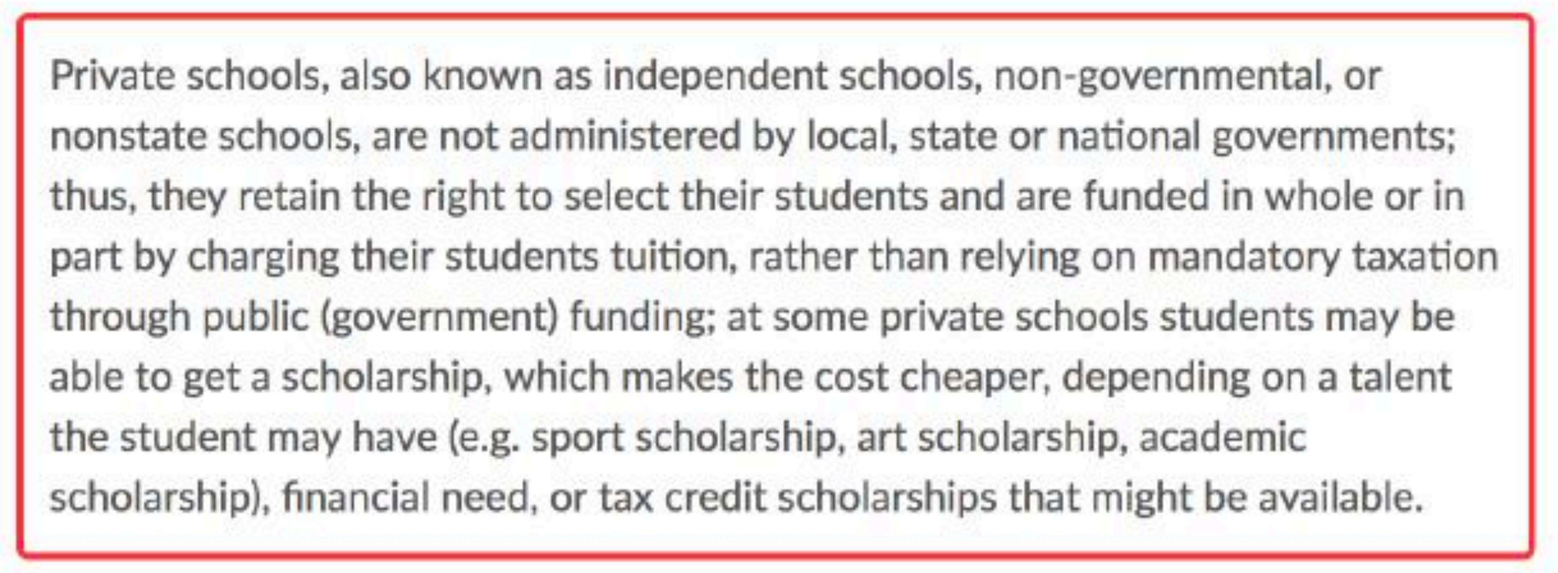
Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. The American Football Conference (AFC) champion Denver Broncos defeated the National Football Conference (NFC) champion Carolina Panthers 24–10 to earn their third Super Bowl title. The game was played on February 7, 2016, at Levi's Stadium in the San Francisco Bay Area at Santa Clara, California.

100k examples

Answer must be a span in the passage

A.k.a. extractive question answering

Stanford Question Answering Dataset (SQuAD)



Private schools, also known as independent schools, non-governmental, or nonstate schools, are not administered by local, state or national governments; thus, they retain the right to select their students and are funded in whole or in part by charging their students tuition, rather than relying on mandatory taxation through public (government) funding; at some private schools students may be able to get a scholarship, which makes the cost cheaper, depending on a talent the student may have (e.g. sport scholarship, art scholarship, academic scholarship), financial need, or tax credit scholarships that might be available.

Along with non-governmental and nonstate schools, what is another name for private schools?

Gold answers: ① independent ② independent schools ③ independent schools

Along with sport and art, what is a type of talent scholarship?

Gold answers: ① academic ② academic ③ academic

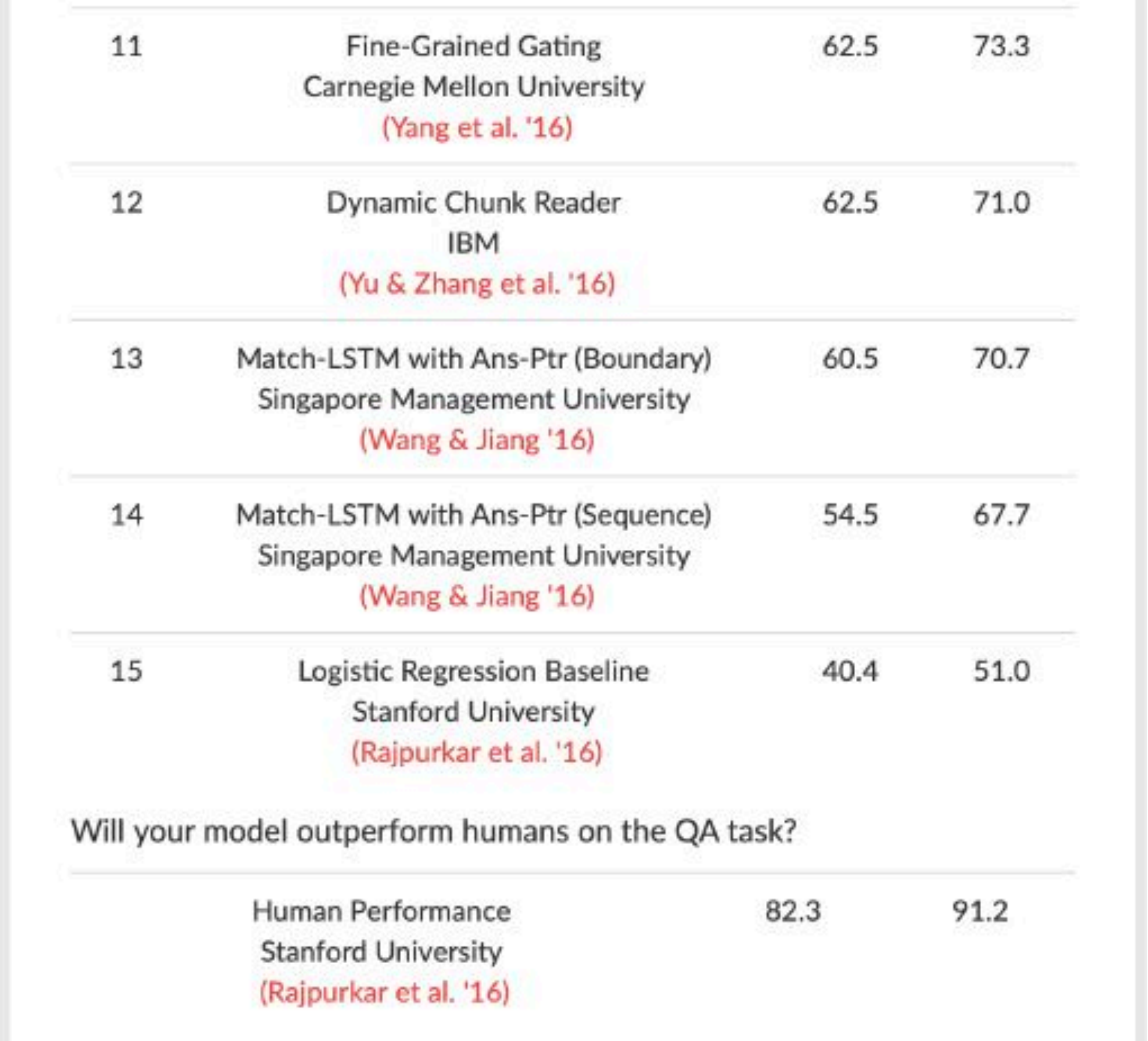
Rather than taxation, what are private schools largely funded by?

Gold answers: ① tuition ② charging their students tuition ③ tuition

SQuAD evaluation, v1.1

- Authors collected 3 gold answers
- Systems are scored on two metrics:
    - Exact match: 1/0 accuracy on whether you match one of the 3 answers
    - F1: Take system and each gold answer as bag of words, evaluate Precision = $\frac{TP}{TP+FP}$, Recall = $\frac{TP}{TP+FN}$, harmonic mean F1 = $\frac{2PR}{P+R}$
      Score is (macro-)average of per-question F1 scores
- F1 measure is seen as more reliable and taken as primary
    - It's less based on choosing exactly the same span that humans chose, which is susceptible to various effects, including line breaks
- Both metrics ignore punctuation and articles (**a**, **an**, **the** only)

SQuAD v1.1 leaderboard, end of 2016 (Dec 6)

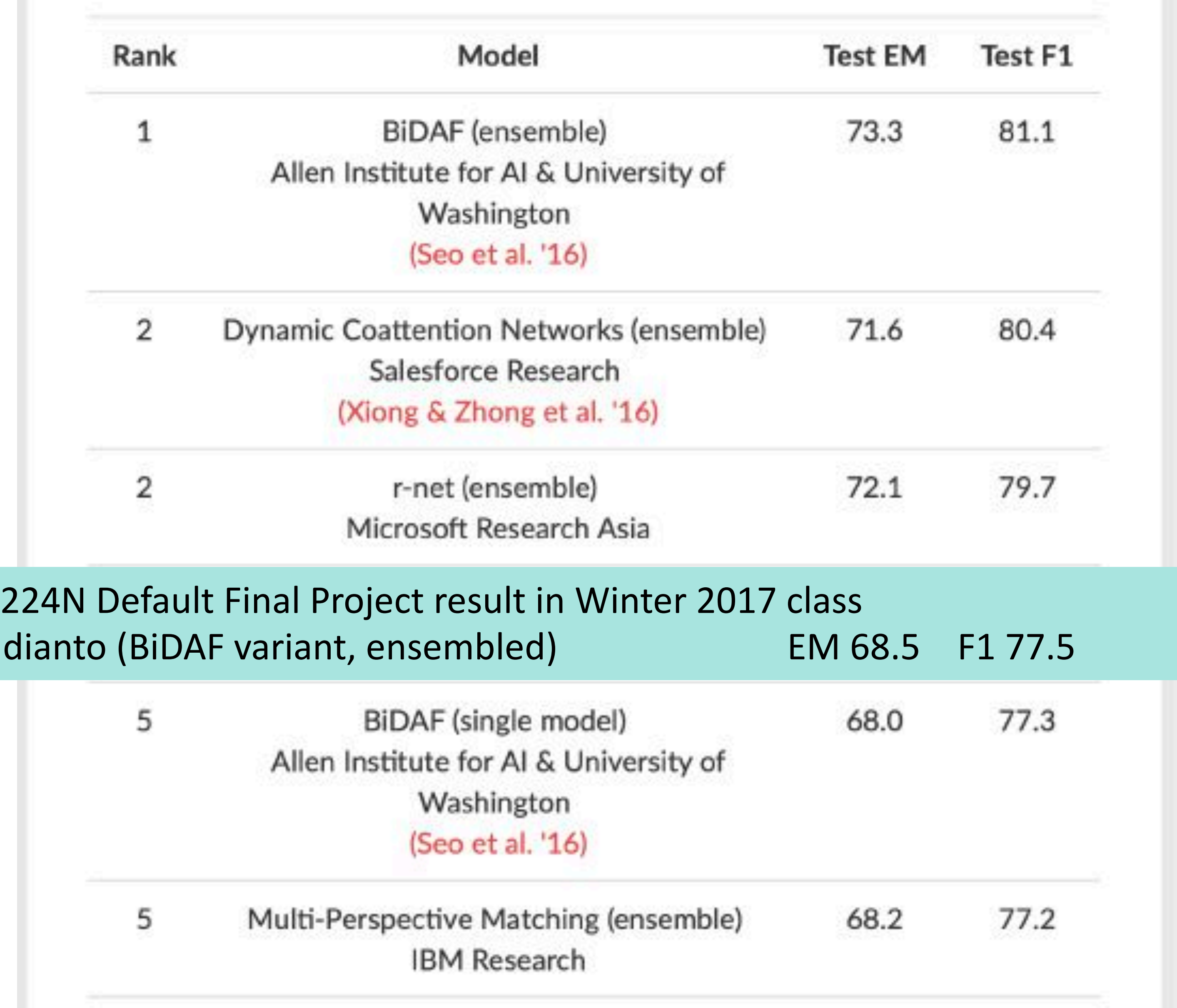


		EM	F1
11	Fine-Grained Gating Carnegie Mellon University (Yang et al. '16)	62.5	73.3
12	Dynamic Chunk Reader IBM (Yu & Zhang et al. '16)	62.5	71.0
13	Match-LSTM with Ans-Ptr (Boundary) Singapore Management University (Wang & Jiang '16)	60.5	70.7
14	Match-LSTM with Ans-Ptr (Sequence) Singapore Management University (Wang & Jiang '16)	54.5	67.7
15	Logistic Regression Baseline Stanford University (Rajpurkar et al. '16)	40.4	51.0

Will your model outperform humans on the QA task?

	Human Performance Stanford University (Rajpurkar et al. '16)	82.3	91.2

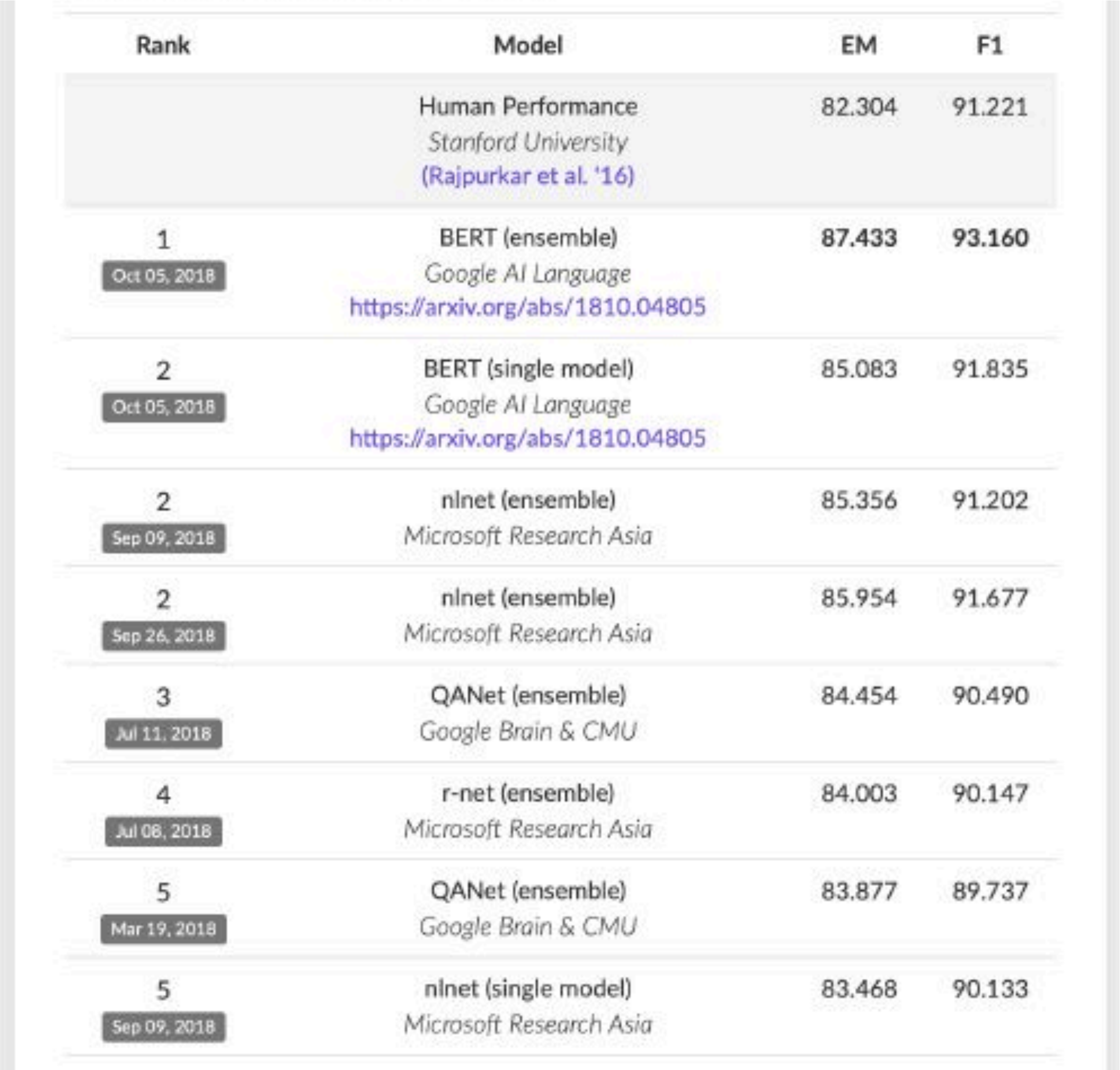
SQuAD v1.1 leaderboard, end of 2016 (Dec 6)



Rank	Model	Test EM	Test F1
1	BiDAF (ensemble) Allen Institute for AI & University of Washington (Seo et al. '16)	73.3	81.1
2	Dynamic Coattention Networks (ensemble) Salesforce Research (Xiong & Zhong et al. '16)	71.6	80.4
2	r-net (ensemble) Microsoft Research Asia	72.1	79.7
5	BiDAF (single model) Allen Institute for AI & University of Washington (Seo et al. '16)	68.0	77.3
5	Multi-Perspective Matching (ensemble) IBM Research	68.2	77.2

Best CS224N Default Final Project result in Winter 2017 class
FNU Budianto (BiDAF variant, ensembled) EM 68.5 F1 77.5

SQuAD v1.1 leaderboard, 2019-02-07 – it's solved!

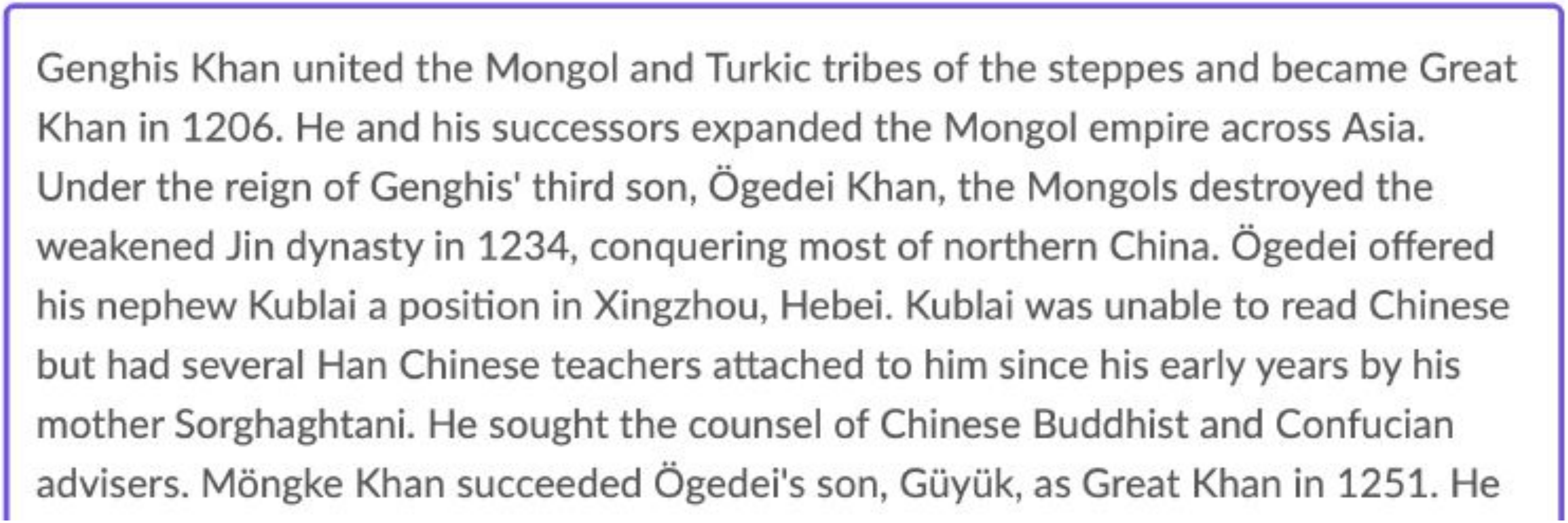


Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Oct 05, 2018	BERT (single model) *Google AI Language* https://arxiv.org/abs/1810.04805	85.083	91.835
2 Sep 09, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.356	91.202
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677
3 Jul 11, 2018	QANet (ensemble) *Google Brain & CMU*	84.454	90.490
4 Jul 08, 2018	r-net (ensemble) *Microsoft Research Asia*	84.003	90.147
5 Mar 19, 2018	QANet (ensemble) *Google Brain & CMU*	83.877	89.737
5 Sep 09, 2018	nlnet (single model) *Microsoft Research Asia*	83.468	90.133

SQuAD 2.0

- A defect of SQuAD 1.0 is that all questions have an answer in the paragraph
- Systems (implicitly) rank candidates and choose the best one
- You don't have to judge whether a span answers the question
- In SQuAD 2.0, 1/3 of the training questions have no answer, and about 1/2 of the dev/test questions have no answer
  - For NoAnswer examples, NoAnswer receives a score of 1, and any other response gets 0, for both exact match and F1
- Simplest system approach to SQuAD 2.0:
  - Have a threshold score for whether a span answers a question
- Or you could have a second component that confirms answering
  - Like Natural Language Inference (NLI) or "Answer validation"

SQuAD 2.0 Example



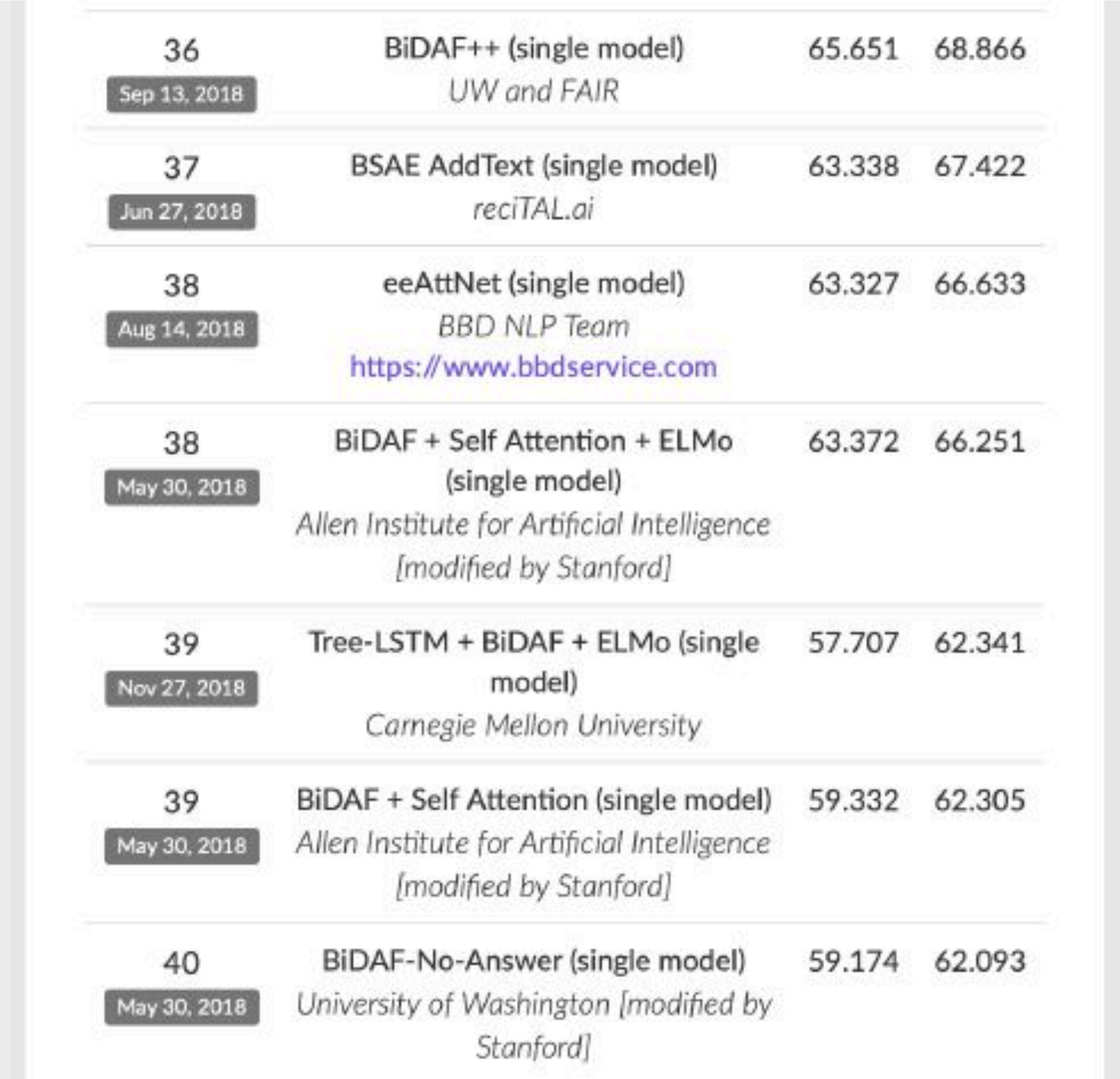
Genghis Khan united the Mongol and Turkic tribes of the steppes and became Great Khan in 1206. He and his successors expanded the Mongol empire across Asia. Under the reign of Genghis' third son, Ögedei Khan, the Mongols destroyed the weakened Jin dynasty in 1234, conquering most of northern China. Ögedei offered his nephew Kublai a position in Xingzhou, Hebei. Kublai was unable to read Chinese but had several Han Chinese teachers attached to him since his early years by his mother Sorghaghtani. He sought the counsel of Chinese Buddhist and Confucian advisers. Möngke Khan succeeded Ögedei's son, Güyük, as Great Khan in 1251. He

When did Genghis Khan kill Great Khan?

Gold Answers: <No Answer>

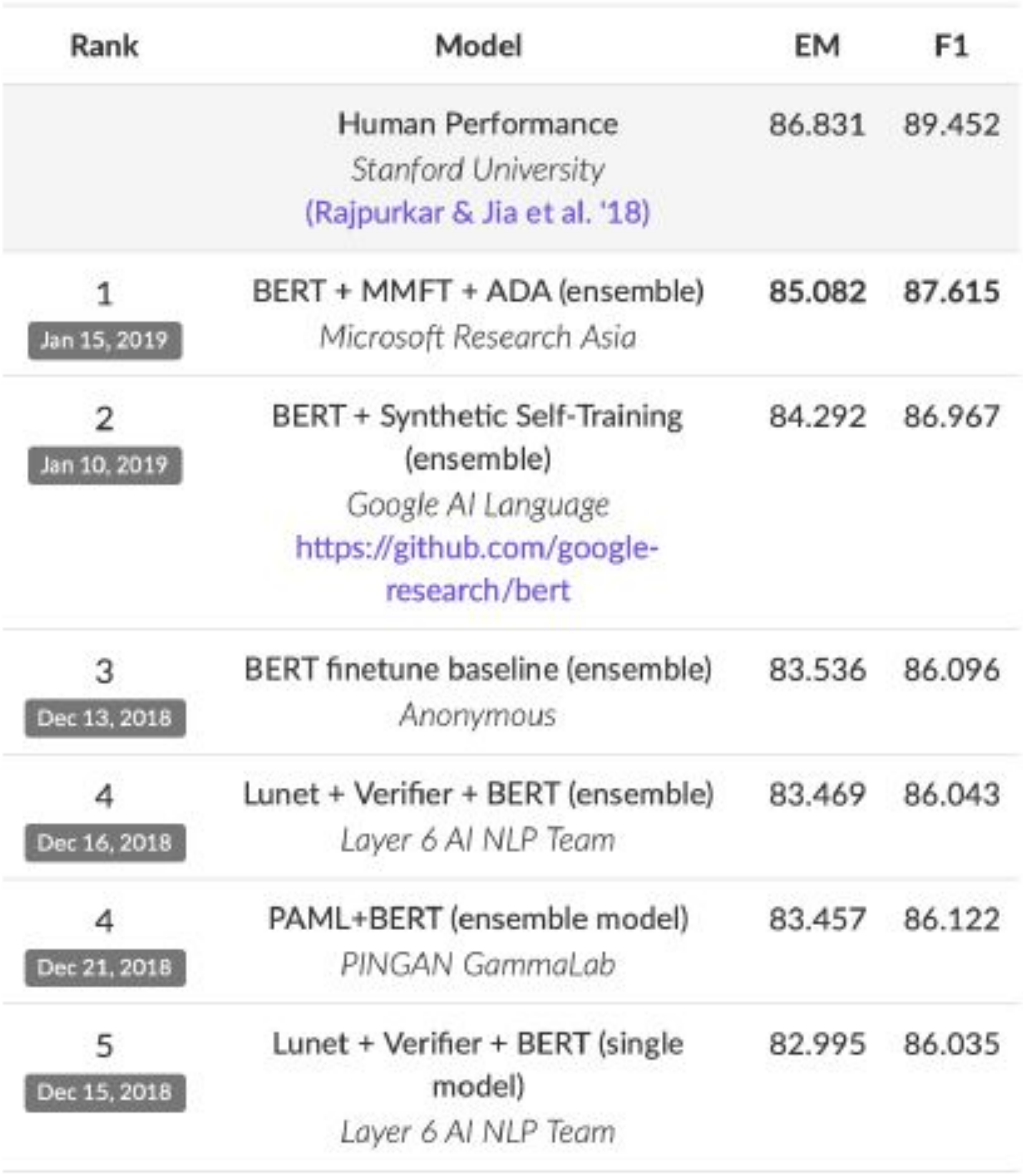
Prediction: 1234 [from Microsoft nlnet]

SQuAD 2.0 leaderboard, 2019-02-07



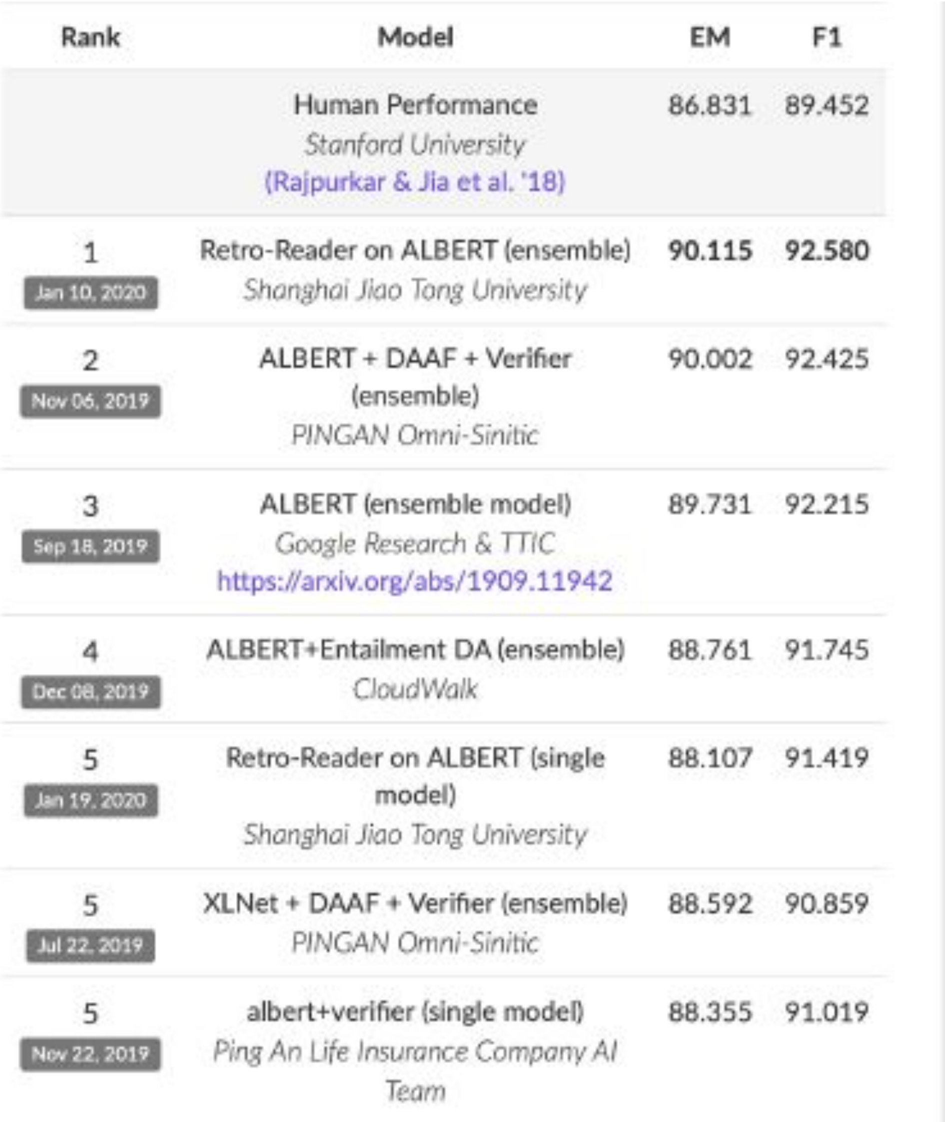
		EM	F1
36 Sep 13, 2018	BiDAF++ (single model) *UW and FAIR*	65.651	68.866
37 Jun 27, 2018	BSAE AddText (single model) *reciTAL.ai*	63.338	67.422
38 Aug 14, 2018	eeAttNet (single model) *BBD NLP Team* https://www.bbdservice.com	63.327	66.633
38 May 30, 2018	BiDAF + Self Attention + ELMo (single model) *Allen Institute for Artificial Intelligence [modified by Stanford]*	63.372	66.251
39 Nov 27, 2018	Tree-LSTM + BiDAF + ELMo (single model) *Carnegie Mellon University*	57.707	62.341
39 May 30, 2018	BiDAF + Self Attention (single model) *Allen Institute for Artificial Intelligence [modified by Stanford]*	59.332	62.305
40 May 30, 2018	BiDAF-No-Answer (single model) *University of Washington [modified by Stanford]*	59.174	62.093

SQuAD 2.0 leaderboard, 2019-02-07



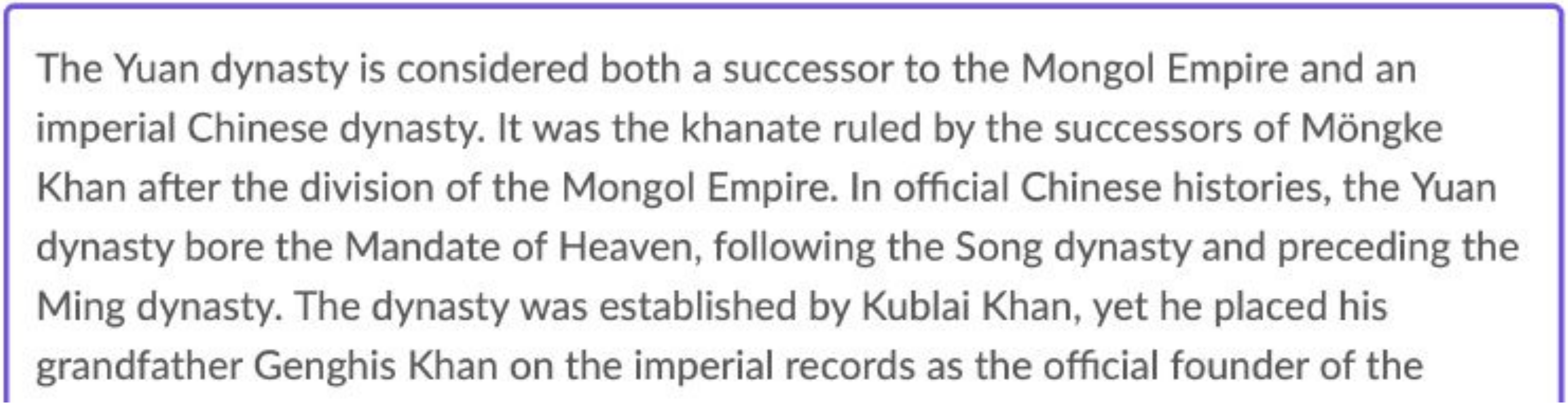
Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 15, 2019	BERT + MMFT + ADA (ensemble) *Microsoft Research Asia*	**85.082**	**87.615**
2 Jan 10, 2019	BERT + Synthetic Self-Training (ensemble) *Google AI Language* https://github.com/google-research/bert	84.292	86.967
3 Dec 13, 2018	BERT finetune baseline (ensemble) *Anonymous*	83.536	86.096
4 Dec 16, 2018	Lunet + Verifier + BERT (ensemble) *Layer 6 AI NLP Team*	83.469	86.043
4 Dec 21, 2018	PAML+BERT (ensemble model) *PINGAN GammaLab*	83.457	86.122
5 Dec 15, 2018	Lunet + Verifier + BERT (single model) *Layer 6 AI NLP Team*	82.995	86.035

SQuAD 2.0 leaderboard, 2020-02-04



Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar & Jia et al. '18)	86.831	89.452
1 Jan 10, 2020	Retro-Reader on ALBERT (ensemble) *Shanghai Jiao Tong University*	**90.115**	**92.580**
2 Nov 06, 2019	ALBERT + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	90.002	92.425
3 Sep 18, 2019	ALBERT (ensemble model) *Google Research & TTIC* https://arxiv.org/abs/1909.11942	89.731	92.215
4 Dec 08, 2019	ALBERT+Entailment DA (ensemble) *CloudWalk*	88.761	91.745
5 Jan 19, 2020	Retro-Reader on ALBERT (single model) *Shanghai Jiao Tong University*	88.107	91.419
5 Jul 22, 2019	XLNet + DAAF + Verifier (ensemble) *PINGAN Omni-Sinitic*	88.592	90.859
5 Nov 22, 2019	albert+verifier (single model) *Ping An Life Insurance Company AI Team*	88.355	91.019

Good systems are great, but still basic NLU errors



The Yuan dynasty is considered both a successor to the Mongol Empire and an imperial Chinese dynasty. It was the khanate ruled by the successors of Möngke Khan after the division of the Mongol Empire. In official Chinese histories, the Yuan dynasty bore the Mandate of Heaven, following the Song dynasty and preceding the Ming dynasty. The dynasty was established by Kublai Khan, yet he placed his grandfather Genghis Khan on the imperial records as the official founder of the

What dynasty came before the Yuan?

Gold Answers: ① Song dynasty ② Mongol Empire
③ the Song dynasty

Prediction: Ming dynasty [BERT (single model) (Google AI)]

SQuAD limitations

- SQuAD has a number of other key limitations too:
  - Only span-based answers (no yes/no, counting, implicit why)
  - Questions were constructed looking at the passages
    - Not genuine information needs
    - Generally greater lexical and syntactic matching between questions and answer span than you get IRL
  - Barely any multi-fact/sentence inference beyond coreference

- Nevertheless, it is a well-targeted, well-structured, clean dataset
  - It has been the most used and competed on QA dataset
  - It has also been a useful starting point for building systems in industry (though in-domain data always really helps!)
  - And we’re using it (SQuAD 2.0)

Good luck with your projects!