Final Project

SQuAD

SQuAD是一个textual question answering数据集。

输入text, question，要求预测answer。

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>!]

Project Proposal

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 and how you plan to extend or improve it in your final project work
   - Suggest a good milestone to have achieved as a halfway point
3. 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

2–4 pages. Details released this Thursday

Due Thu Feb 14, 4:30pm on Gradescope

Project Milestoney

- 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)

Find SOTA

link[http://paperswithcode.com/sota]

Dataset

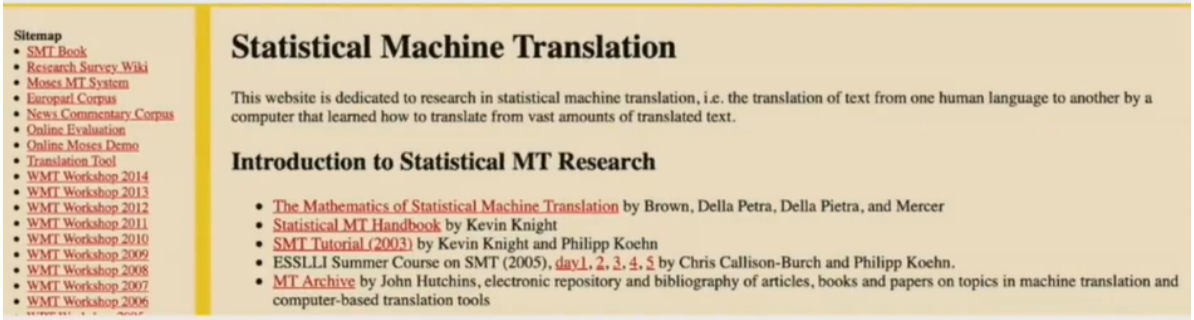
Linguistic Data Consortium

- 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!

24

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

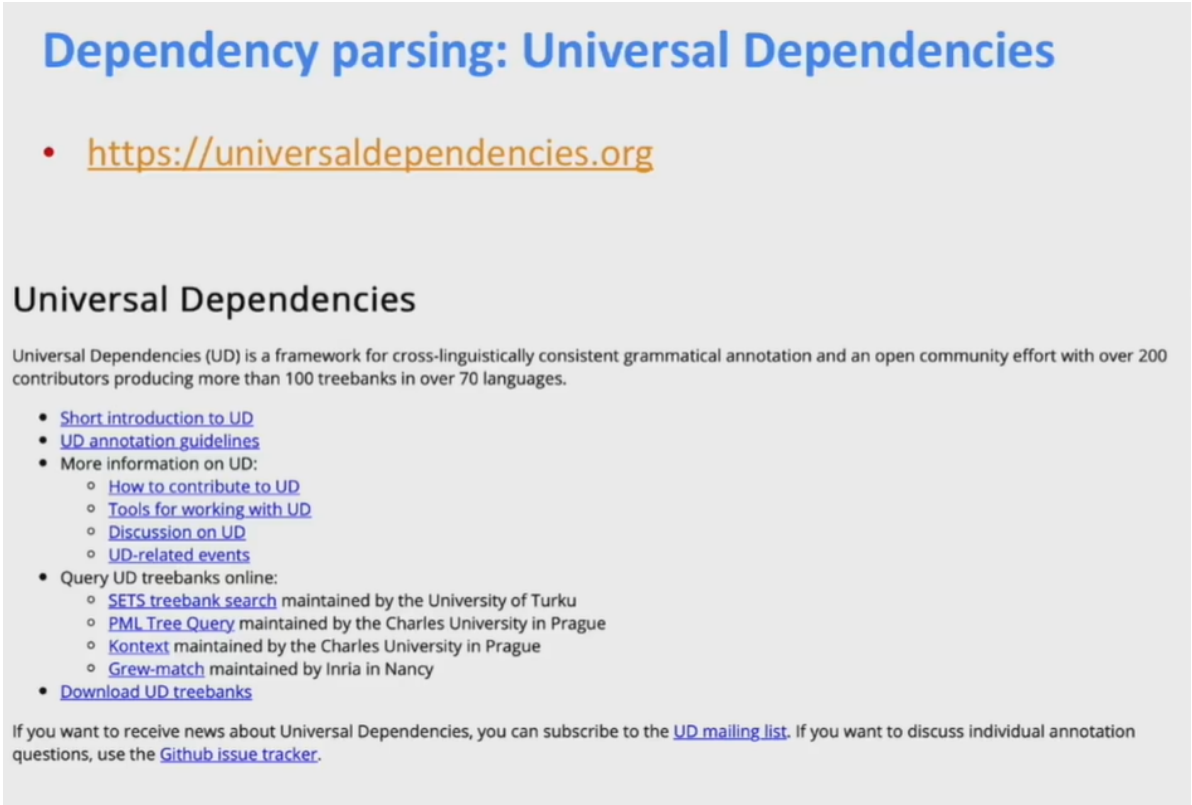
This website is dedicated to research in statistical machine translation, i.e. the translation of text from one human language to another by a computer that learned how to translate from vast amounts of translated text.

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.

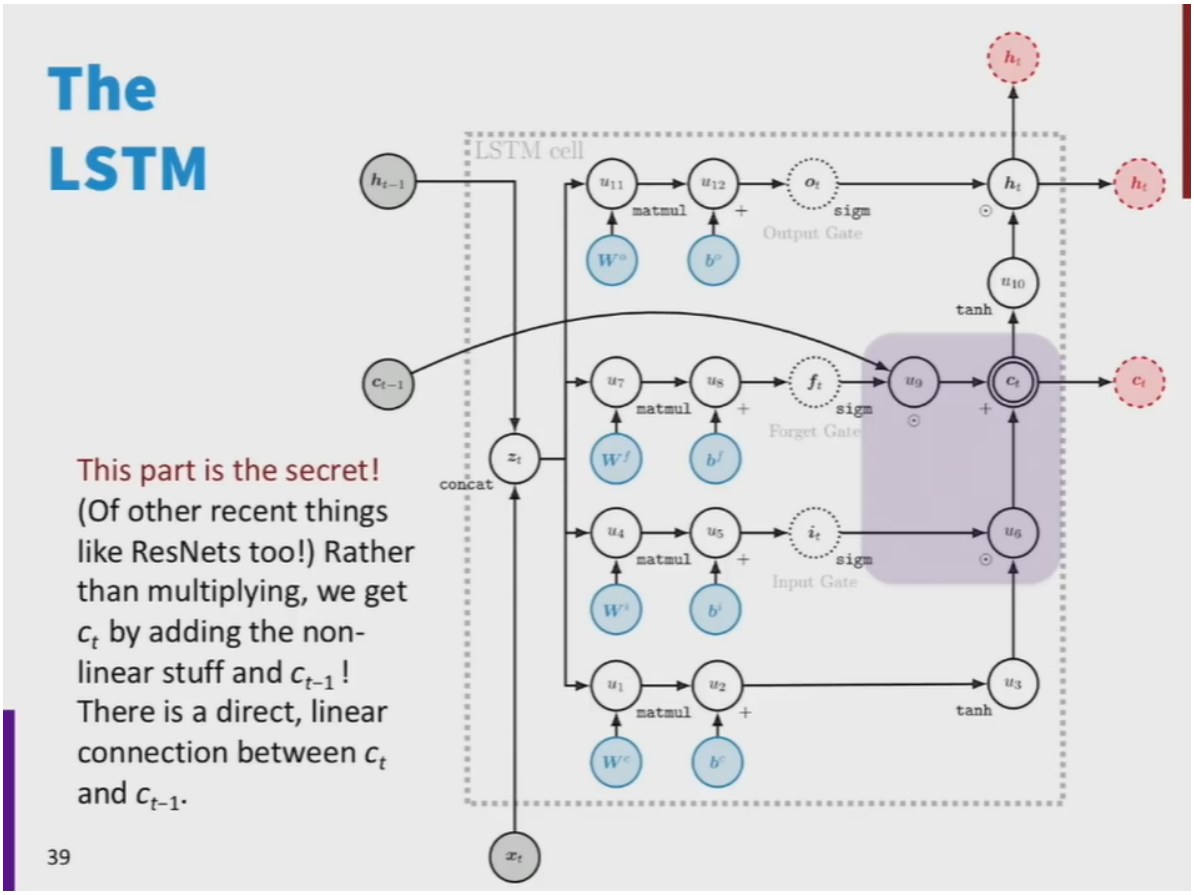
- Short introduction to UD
- UD annotation guidelines
- More information on UD:
  - How to contribute to UD
  - Tools for working with UD
  - Discussion on UD
  - UD-related events
- Query UD treebanks online:
  - SETS treebank search maintained by the University of Turku
  - PML Tree Query maintained 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
  - Ask on Piazza or talk to course staff

LSTM Review



h的计算用到了残差的思想。

Large output is a problem

NMT中每个timestep要出一个单词，需要一个$|V| * n$的权重矩阵，来和h作乘法，最后softmax。

这个权重矩阵计算消耗很大。

UNK is a problem

UNK指语料库没有的词，无法翻译。

解决方法(和上面large output一起)

- *Hierarchical softmax*: tree-structured vocabulary
- *Noise-contrastive estimation*: binary classification
- *Train* on a subset of the vocabulary at a time; *test* on a smart on the set of possible translations
  - *Jean, Cho, Memisevic, Bengio. ACL2015*
- *Use attention to work out what you are translating:* You can do something simple like dictionary lookup
- *More ideas we will get to:* Word pieces; char. models

MT Evaluation

- Manual (the best!?):
  - **Adequacy and Fluency** (5 or 7 point scales)
  - Error categorization
  - **Comparative ranking of translations**
- Testing in an application that uses MT as one sub-component
  - E.g., question answering from foreign language documents
    - May not test many aspects of the translation (e.g., cross-lingual IR)
- Automatic metric:
  - **BLEU (Bilingual Evaluation Understudy)**
  - Others like TER, METEOR, …

BLEU

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

Reference (human) translation:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

Machine translation:
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

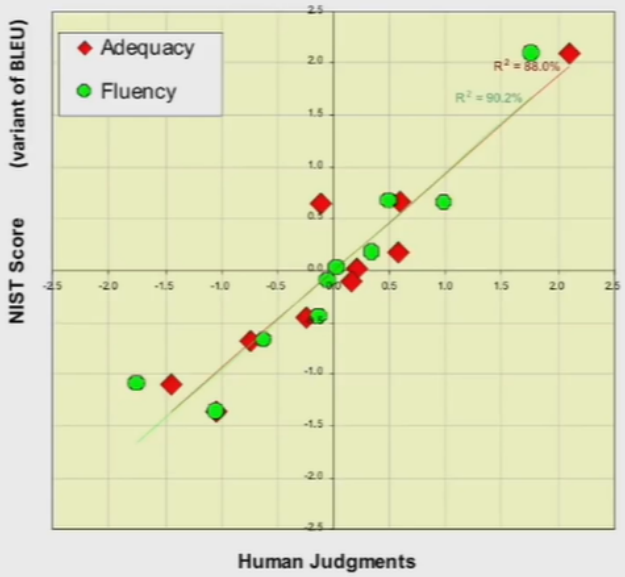
- BLEU is a weighted geometric mean, with a brevity penalty factor added.
  - Note that it's precision-oriented
- BLEU4 formula
  (counts n-grams up to length 4)

exp (0.5 * log p1 +
0.25 * log p2 +
0.125 * log p3 +
0.125 * log p4 –
max(words-in-reference / words-in-machine – 1, 0)

p1 = 1-gram precision
P2 = 2-gram precision
P3 = 3-gram precision
P4 = 4-gram precision

Note: only works at corpus level (zeroes kill it); there's a smoothed variant for sentence-level

Initial results showed that BLEU predicts human judgments well



Automatic evaluation of MT

- People started optimizing their systems to maximize BLEU score
  - BLEU scores improved rapidly
  - The correlation between BLEU and human judgments of quality went way, way down
  - MT BLEU scores now approach those of human translations but their true quality remains far below human translations
- Coming up with automatic MT evaluations has become its own research field
  - There are many proposals: TER, METEOR, MaxSim, SEPIA, our own RTE-MT
  - TERpA is a representative good one that handles some word choice variation.
- MT research **requires** *some* automatic metric to allow a rapid development and evaluation cycle.