CS287: Statistical Natural Language Processing

Alexander Rush

April 6, 2016

Contents

Count-based Language Models

By the chain rule, any distribution can be factorized as

$$p(w_1, \ldots, w_T) = \prod_{t=1}^{T} p(w_t | w_1, \ldots, w_{t-1})$$

Count-based n-gram language models make a Markov assumption:

$$p(w_t | w_1, \ldots, w_t) \approx p(w_t | w_{t-n}, \ldots, w_{t-1})$$

Need smoothing to deal with rare n-grams.

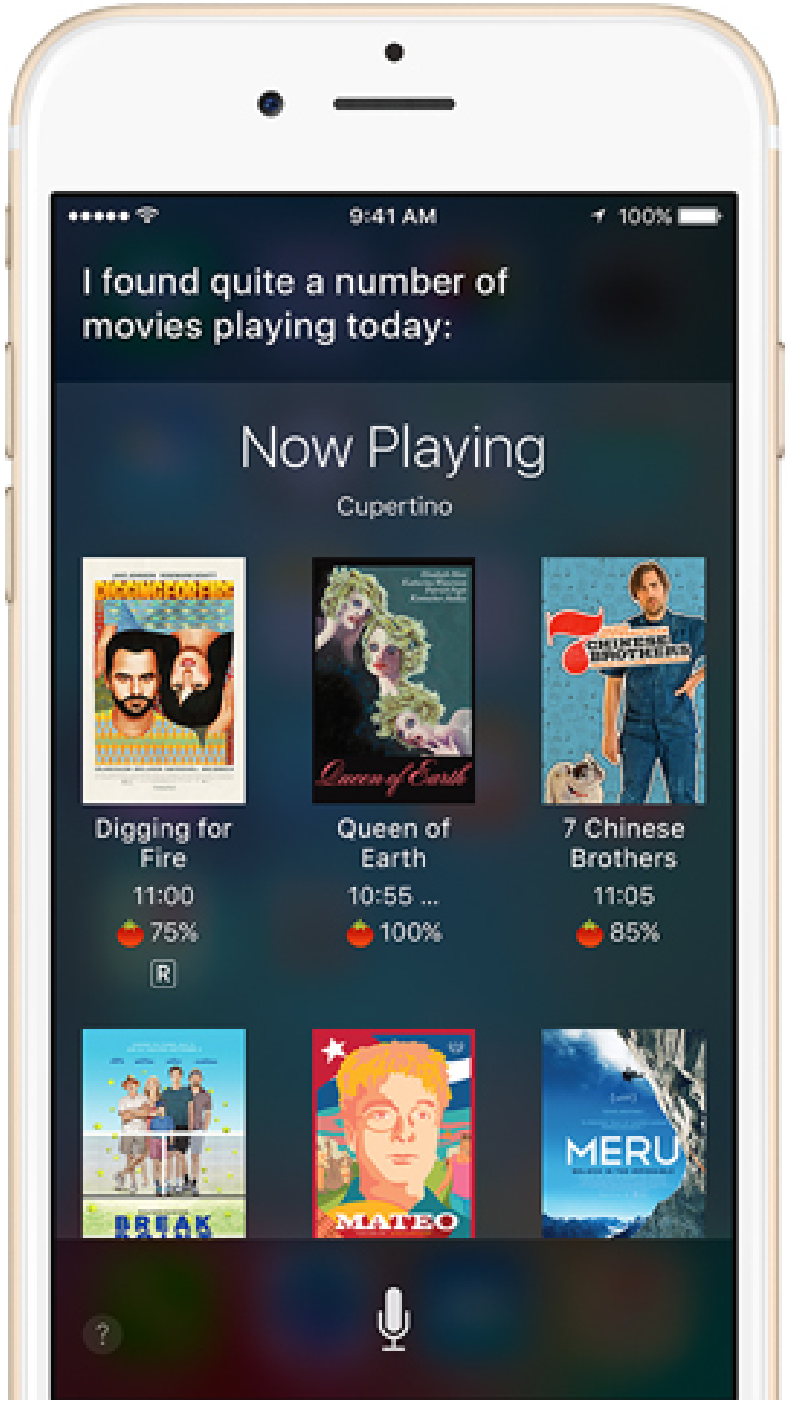
Neural Language Models

Neural Language Models (NLM)

- Represent words as dense vectors in $\mathbb{R}^n$ (word embeddings).

  $\mathbf{w}_t \in \mathbb{R}^{|\mathcal{V}|}$: One-hot representation of word $\in \mathcal{V}$ at time t
  $\Rightarrow \mathbf{x}_t = \mathbf{X}\mathbf{w}_t$: Word embedding ($\mathbf{X} \in \mathbb{R}^{n \times |\mathcal{V}|}, n < |\mathcal{V}|$)

- Train a neural net that composes history to predict next word.


9:41 AM
100%
I found quite a number of movies playing today:
Now Playing
Cupertino
Digging for Fire
11:00
75%
R
Queen of Earth
10:55 ...
100%
7 Chinese Brothers
11:05
85%
BREAK
MATEO
MERU


amazon

Abraham Lincoln

16th U.S. President

Abraham Lincoln was the 16th President of the United States, serving from March 1861 until his assassination in April 1865. Wikipedia

Born: February 12, 1809, Hodgenville, KY

Height: 6′ 4″

Spouse: Mary Todd Lincoln (m. 1842–1865)

Party: National Union Party

Children: William Wallace Lincoln, Robert Todd Lincoln, Tad Lincoln, Edward Baker Lincoln

Quotes

View 7+ more

Nearly all men can stand adversity, but if you want to test a man's character, give him power.

Whatever you are, be a good one.

Always bear in mind that your own resolution to succeed is more important than any other.

People also search for

View 15+ more


George Washington


William Wallace Lincoln
Son


John Wilkes Booth


John F. Kennedy


Mary Todd Lincoln
Spouse

Contents

Foundational Challenge: Turing Test

Q: Please write me a sonnet on the subject of the Forth Bridge.

A : Count me out on this one. I never could write poetry.

Q: Add 34957 to 70764.

A: (Pause about 30 seconds and then give as answer) 105621.

Q: Do you play chess?

A: Yes.

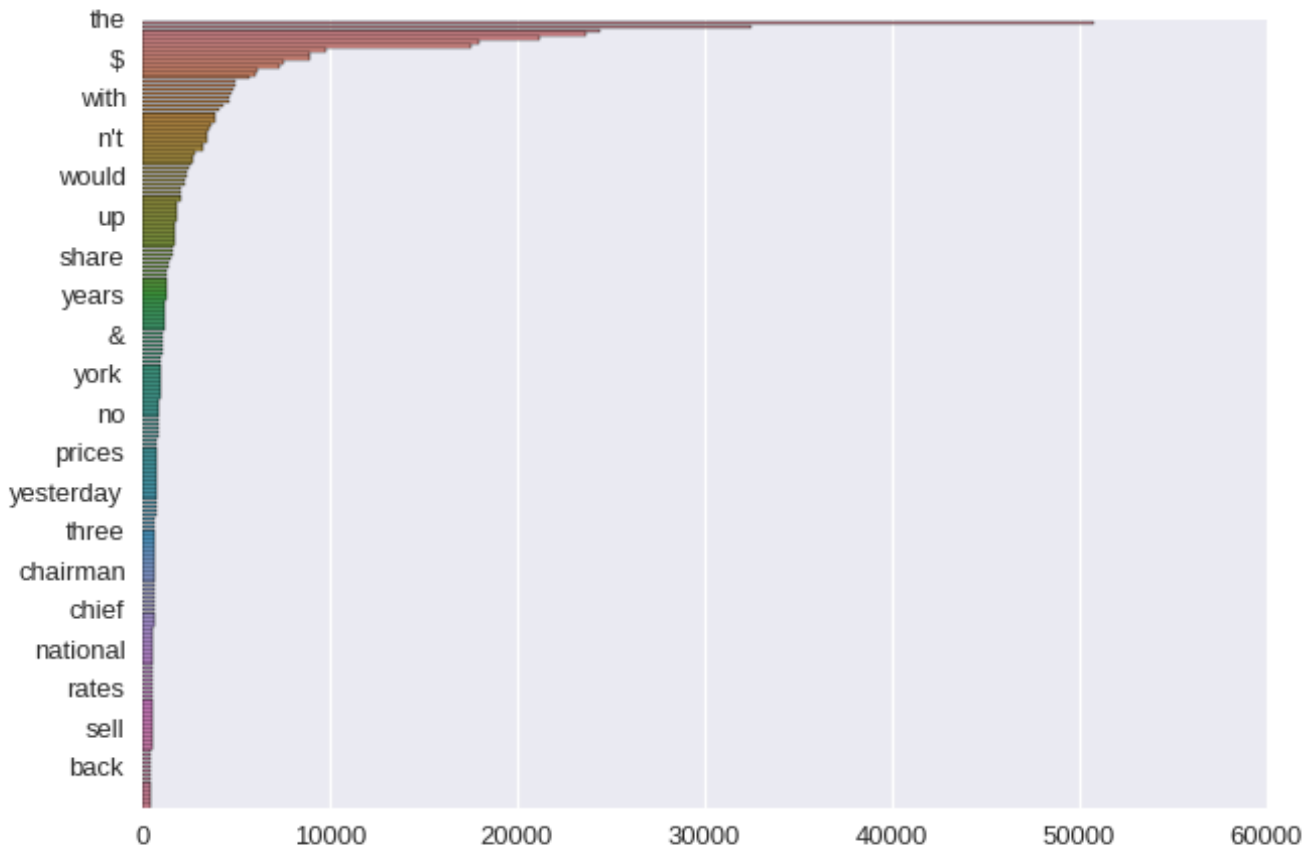
Q: I have K at my K1, and no other pieces. You have only K at K6 and R at R1. It is your move. What do you play?

A: (After a pause of 15 seconds) R-R8 mate. - Turing (1950)

(1) Lexicons and Lexical Semantics

Zipf' Law (1935,1949):

The frequency of any word is inversely proportional to its rank in the frequency table.



(2) Structure and Probabilistic Modeling

The Shannon Game (Shannon and Weaver, 1949):

Given the last n words, can we predict the next one?

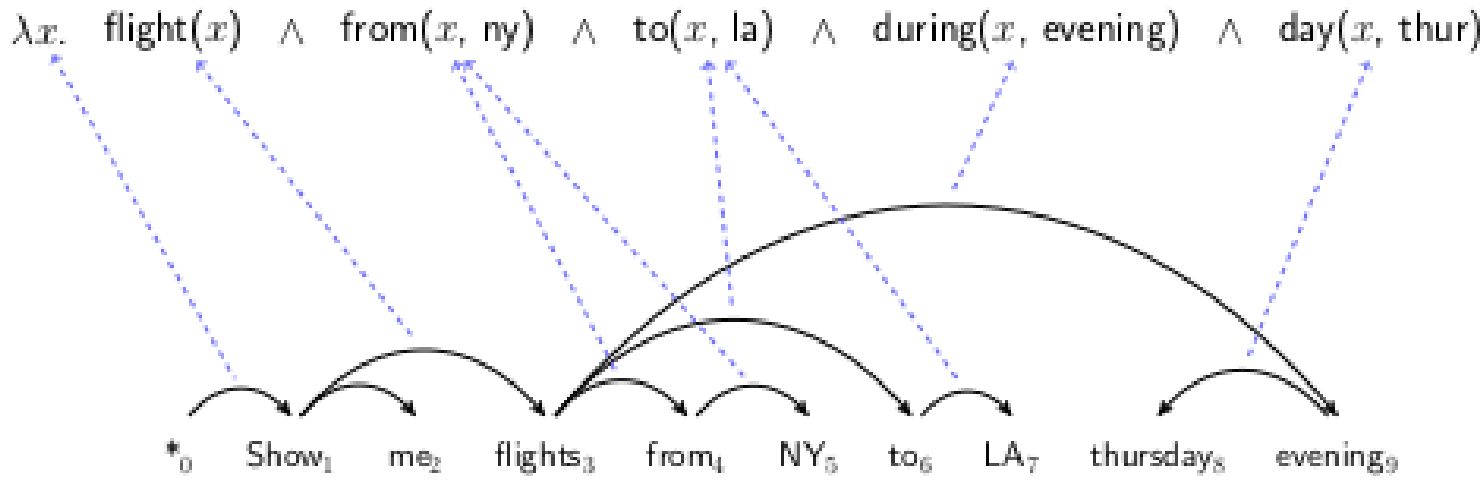
`The pin-tailed snipe (Gallinago stenura) is a small stocky wader. It breeds in northern Russia and migrates to spend the __`

- Probabilistic models have become very effective at this task.
- Crucial for speech recognition (Jelinek), OCR, automatic translations, etc.

(3) Compositionality of Syntax and Semantics

Probabilistic models give no insight into some of the basic problems of syntactic structure - Chomsky (1956)

Show me flights from NY to LA Thursday evening.



(4) Document Structure and Discourse

Language is not merely a bag-of-words but a tool with particular properties - Harris (1954)

(5) Knowledge and Reasoning Beyond the Text

> *It is based on the belief that in modeling language understanding, we must deal in an integrated way with all of the aspects of language syntax, semantics, and inference.* - Winograd (1972)

`The city councilmen refused the demonstrators a permit because they [feared/advocated] violence.`

- Recently (2011) posed as a challenge for testing commonsense reasoning.

Contents

Deep Learning and NLP

- Presentation-based on Chris Manning's "Computational Linguistics and Deep Learning" (2016) published in *Computational Linguistics*

  Deep Learning waved have lapped at the shores of computational linguistics for several years now, but 2015 seems like the year when the full force of the tsunami hit major NLP conferences. - Chris Manning

NLP as a Challenge for Machine Learning

I'd use the billion dollars to build a NASA-size program focusing on natural langauge processing in all of its glory (semantics, pragmatics, etc.) ... Intellectually I think that NLP is fascinating, allowing us to focus on highly structure inference programs, on issues that go to the core of 'what is thought' but remain eminently practical, and on a technology that surely would make the world a better place" - Jordan (2014)

NLP as a Challenge for Deep Learning

The next big step for Deep Learning is natural language understanding, which aims to give machines the power to understand not just individual words but entire sentence and paragraphs. - Bengio

What are they referring to?

Recent advances in,

- Speech Recognition
- Language Modeling
- Machine Translation
- Question Answering
- many other tasks.

Still,

> *Problems in higher-level language processing have not seen the dramatic error-rate reductions from deep learning that have been seen in speech recognition and object recognition in vision.*

What are they referring to?

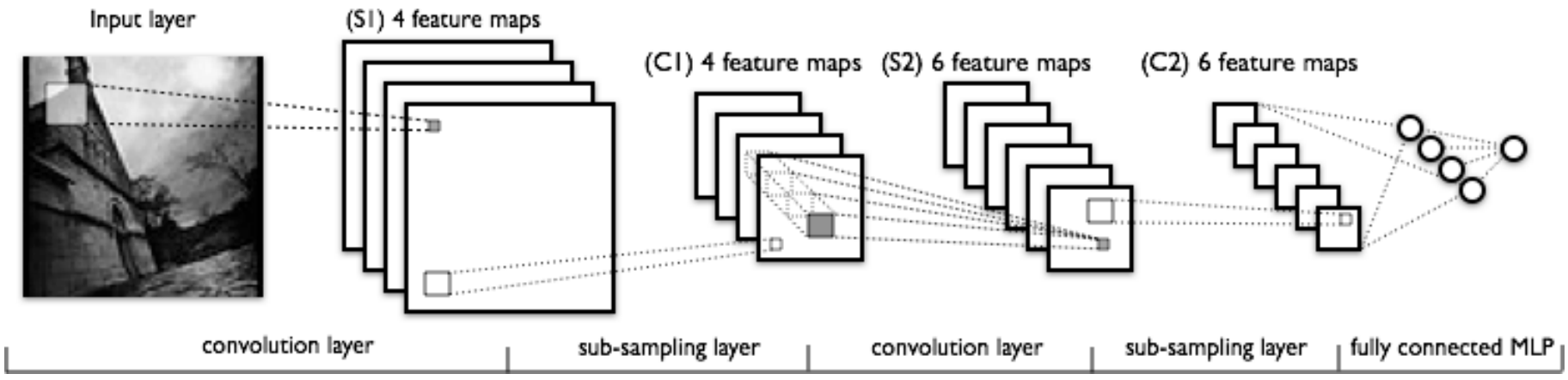
Recent advances in,

- Speech Recognition
- Language Modeling
- Machine Translation
- Question Answering
- many other tasks.

Still,

> *Problems in higher-level language processing have not seen the dramatic error-rate reductions from deep learning that have been seen in speech recognition and object recognition in vision.*

Object Recognition



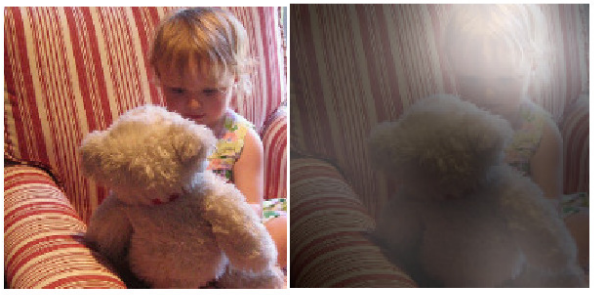

Layer 1
Layer 2
Layer 3
Layer 4
Layer 5

Image Captioning

A woman is throwing a frisbee in a park.

A dog is standing on a hardwood floor.

A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.

A group of people sitting on a boat in the water.

A giraffe standing in a forest with trees in the background.

Central Aspects of Deep Learning for NLP

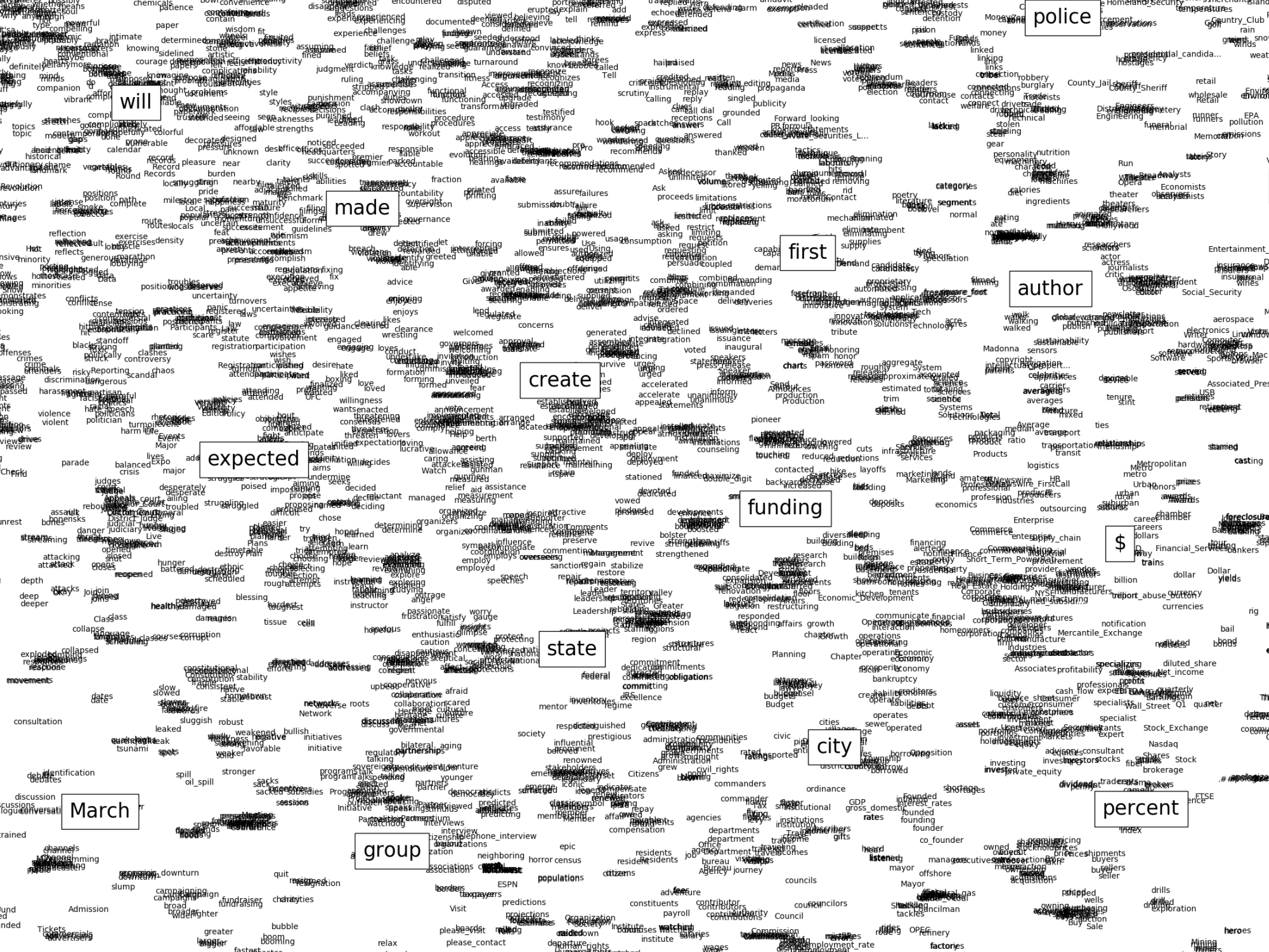
1. Learn the features representations of language.
2. Construct higher-level structure in a latent manner
3. Train systems completely end-to-end.

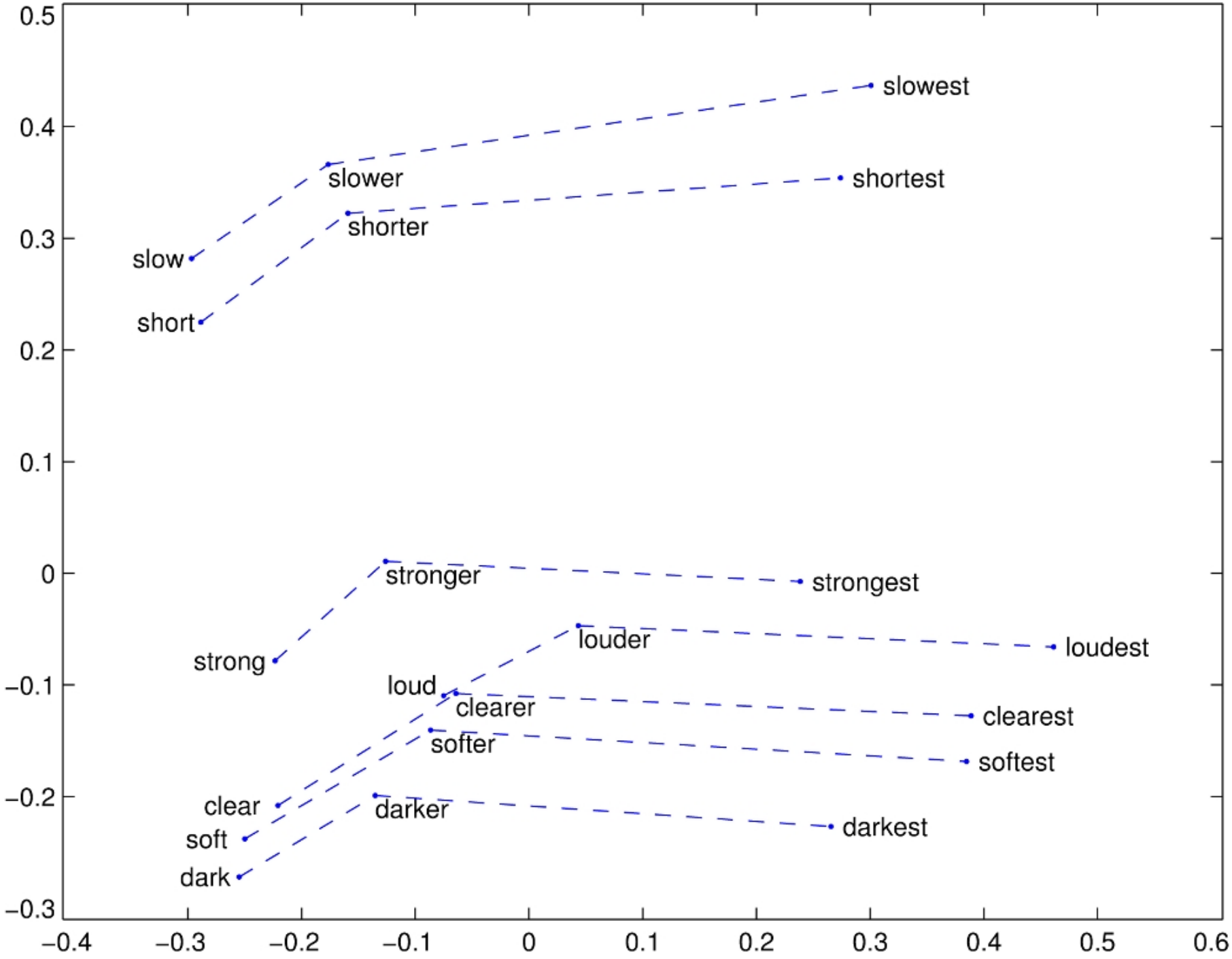
Central Aspects of Deep Learning for NLP

1. Learn the features representations of language.
2. Construct higher-level structure in a latent manner
3. Train systems completely end-to-end.

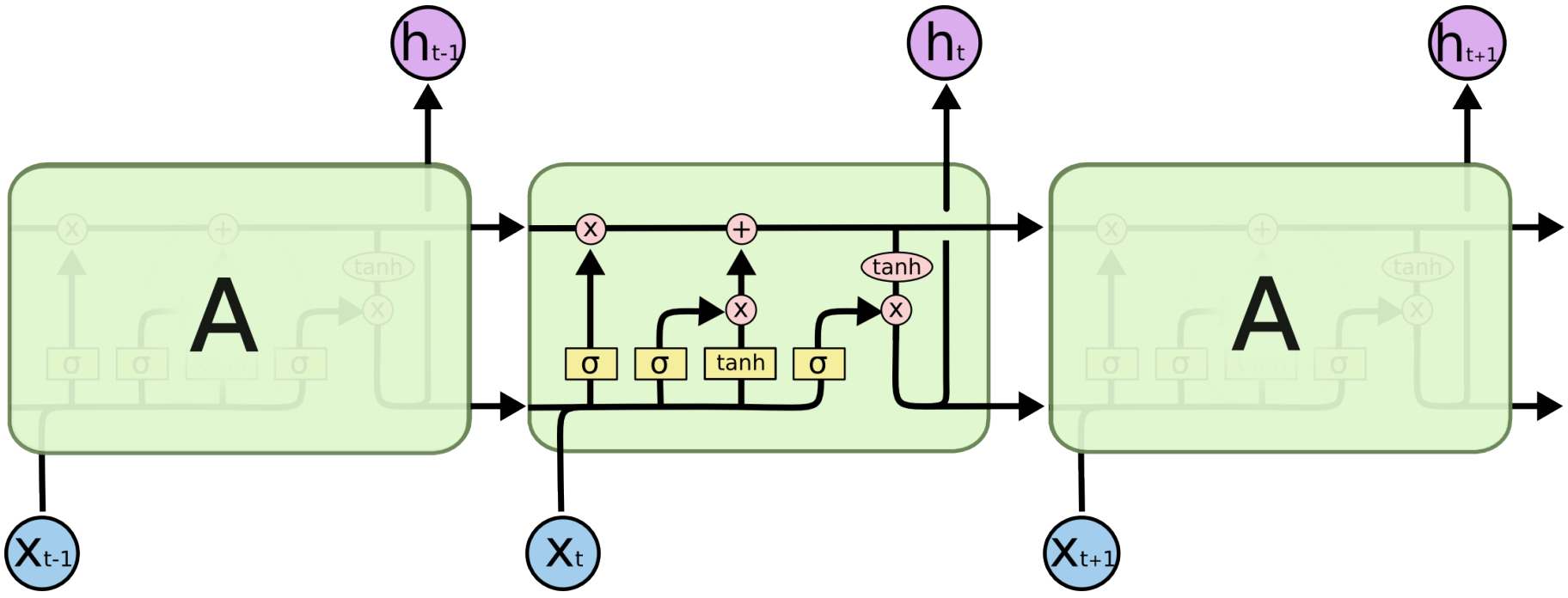
Central Aspects of Deep Learning for NLP

1. Learn the features representations of language.
2. Construct higher-level structure in a latent manner
3. Train systems completely end-to-end.


will
made
police
first
author
create
expected
funding
$
state
city
March
percent
group


0.5
0.4
0.3
0.2
0.1
0
−0.1
−0.2
−0.3
−0.4
−0.3
−0.2
−0.1
0
0.1
0.2
0.3
0.4
0.5
0.6
slowest
slower
shortest
shorter
slow
short
stronger
strongest
louder
loudest
strong
loud
clearer
clearest
softer
softest
clear
darker
darkest
soft
dark

LSTM



Proof. Omitted. □

Lemma 0.1. *Let $\mathcal{C}$ be a set of the construction.*

Let $\mathcal{C}$ be a gerber covering. Let $\mathcal{F}$ be a quasi-coherent sheaves of $\mathcal{O}$-modules. We have to show that

$$\mathcal{O}_{\mathcal{O}_X} = \mathcal{O}_X(\mathcal{L})$$

.

Proof. This is an algebraic space with the composition of sheaves $\mathcal{F}$ on $X_{\acute{e}tale}$ we have

$$\mathcal{O}_X(\mathcal{F}) = \{morph_1 \times_{\mathcal{O}_X} (\mathcal{G}, \mathcal{F})\}$$

where $\mathcal{G}$ defines an isomorphism $\mathcal{F} \to \mathcal{F}$ of $\mathcal{O}$-modules. □

Lemma 0.2. *This is an integer $\mathcal{Z}$ is injective.*

Proof. See Spaces, Lemma **??**. □

Lemma 0.3. *Let S be a scheme. Let X be a scheme and X is an affine open covering. Let $\mathcal{U} \subset \mathcal{X}$ be a canonical and locally of finite type. Let X be a scheme. Let X be a scheme which is equal to the formal complex.*

The following to the construction of the lemma follows.

Let X be a scheme. Let X be a scheme covering. Let

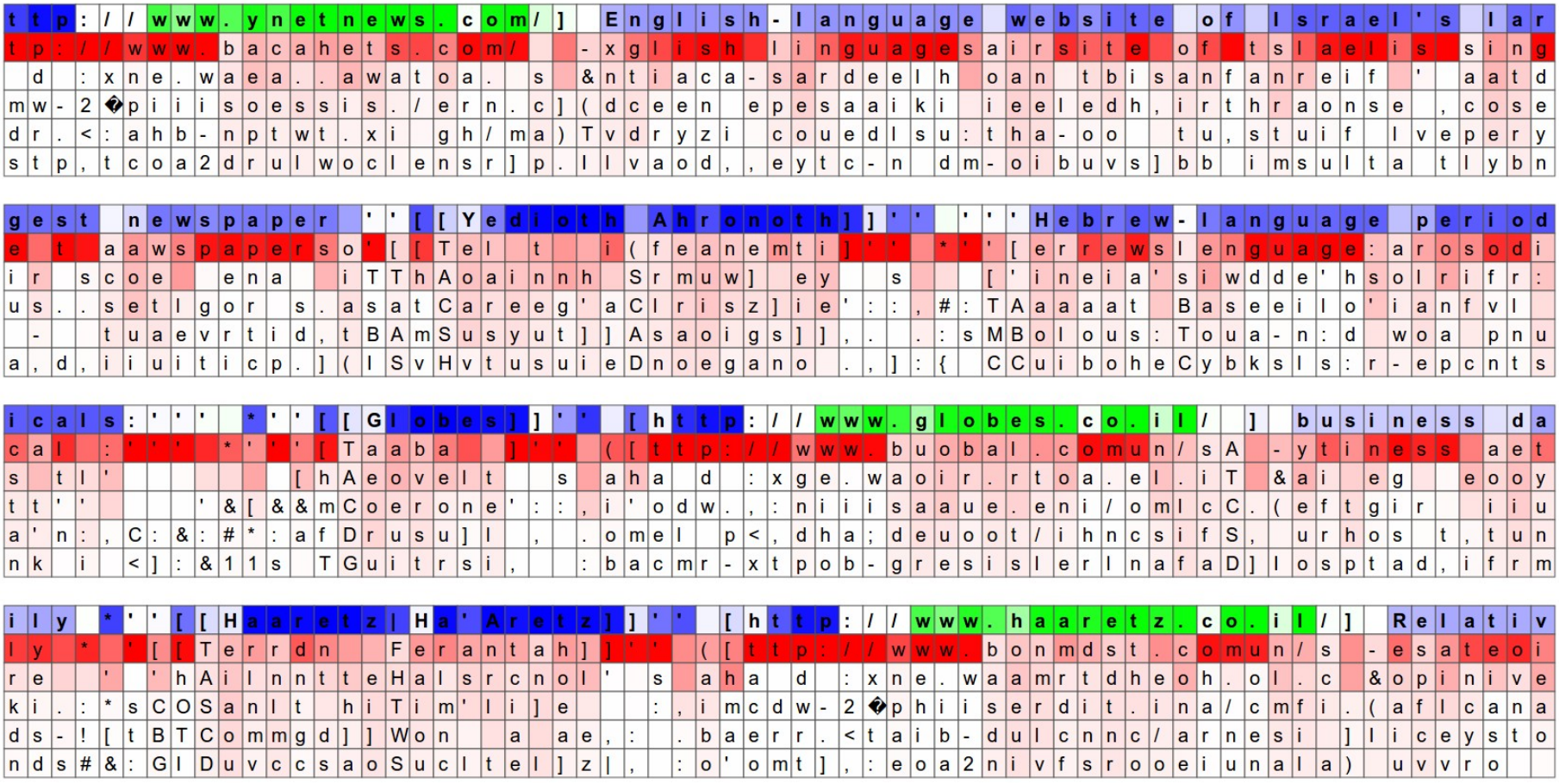
$$b : X \to Y' \to Y \to Y \to Y' \times_X Y \to X.$$

be a morphism of algebraic spaces over S and Y.

Proof. Let X be a nonzero scheme of X. Let X be an algebraic space. Let $\mathcal{F}$ be a quasi-coherent sheaf of $\mathcal{O}_X$-modules. The following are equivalent

(1) $\mathcal{F}$ is an algebraic space over S.
(2) If X is an affine open covering.

Consider a common structure on X and X the functor $\mathcal{O}_X(U)$ which is locally of finite type. □



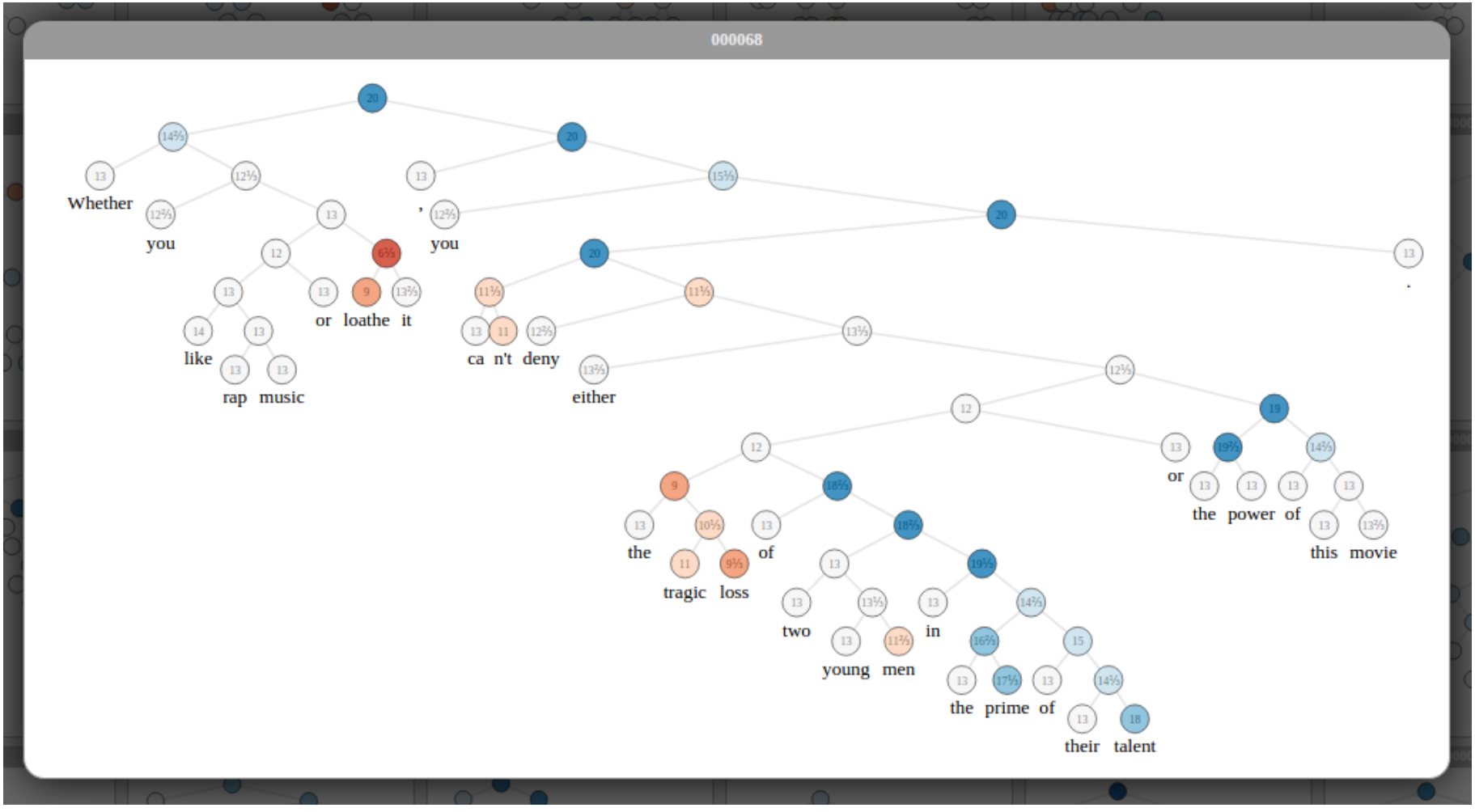
GPU Processing

- Neural Networks are remarkably parallel-izable.

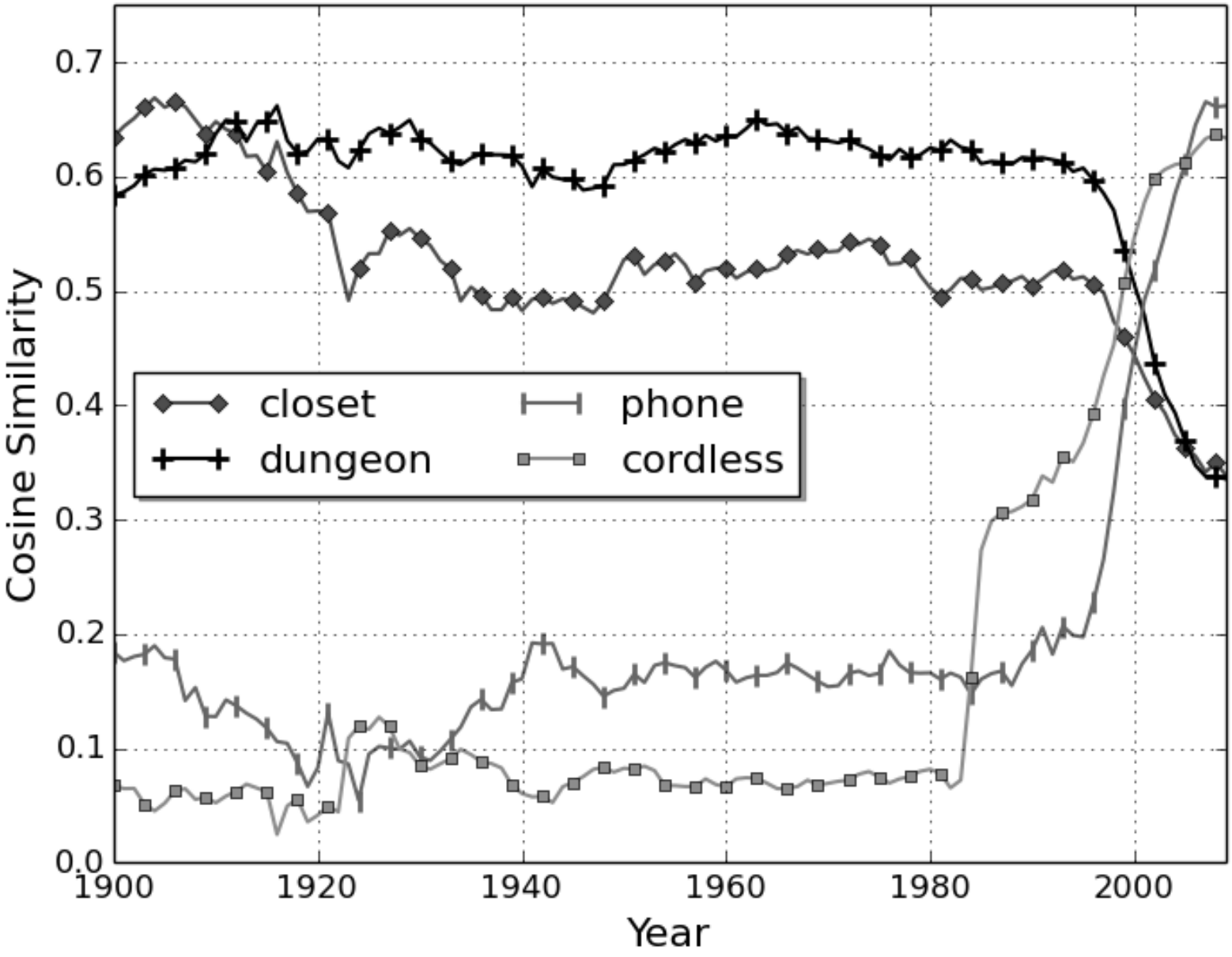
GPU Implementation of a variant of HW1

	non-GPU	GPU
per epoch	2475s	54.0 s
per batch	787ms	15.6 ms

(1) Compositional Structures?



(2) Understanding Text?



(3) Language and Thought?

- Do these methods tell us anything about core nature of language?
- Do they inform psychology or cognitive problems?

Contents

This Semester

Deep Learning for Natural Language Processing

- Primarily a lecture course.
- Topics and papers distributed throughout.
- Main Goal: Educate researchers in NLP

Background

- Some college-level Machine Learning course
- Practical programming experience
- Interest in applied experimental research (not a theory course)

Audience

Take this class to...

- understand about cutting-edge methods in the area.
- replicate many important recent results
- apply machine learning to relevant, interesting problems

Do not take this class to...

- get experience with common NLP tools (NLTK, CoreNLP, etc.)
- build a system for your (non-NLP) startup
- learn much about modern Linguistics

Audience

Take this class to...

- understand about cutting-edge methods in the area.
- replicate many important recent results
- apply machine learning to relevant, interesting problems

Do not take this class to...

- get experience with common NLP tools (NLTK, CoreNLP, etc.)
- build a system for your (non-NLP) startup
- learn much about modern Linguistics

Topics

1. Machine Learning for Text
2. Feed-Forward Neural Networks
3. Language Modeling and Word Embeddings
4. Recurrent Neural Networks
5. Conditional Random Fields and Structured Prediction

Topics

1. Machine Learning for Text
2. Feed-Forward Neural Networks
3. Language Modeling and Word Embeddings
4. Recurrent Neural Networks
5. Conditional Random Fields and Structured Prediction

Topics

1. Machine Learning for Text
2. Feed-Forward Neural Networks
3. Language Modeling and Word Embeddings
4. Recurrent Neural Networks
5. Conditional Random Fields and Structured Prediction

Topics

1. Machine Learning for Text
2. Feed-Forward Neural Networks
3. Language Modeling and Word Embeddings
4. Recurrent Neural Networks
5. Conditional Random Fields and Structured Prediction

Topics

1. Machine Learning for Text
2. Feed-Forward Neural Networks
3. Language Modeling and Word Embeddings
4. Recurrent Neural Networks
5. Conditional Random Fields and Structured Prediction

Homeworks

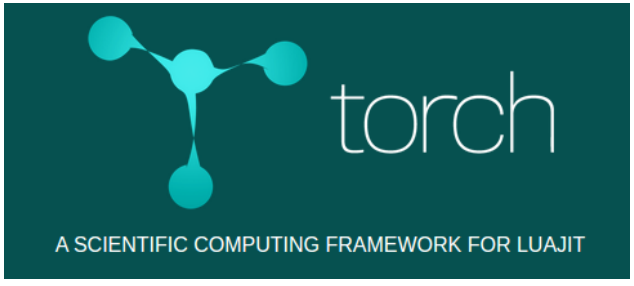
Each homework will require you to replicate a research result,

- Text Classification
- Sentence Tagging
- Language Modeling (1)
- Language Modeling (2) (LSTMs)
- Name-Entity Recognition (CRFs)

Programming

Assignments use,

- Python for text processing and visualization
- Lua/Torch for neural networks



- First section on Fri. will be introduction.

Applications

Lectures on NLP applications

- Language Modeling
- Coreference and Pronoun Anaphora
- Neural Machine Translation
- Syntactic Parsing

Final Project

- Empirical project done in teams
- Research-level project on current topics
- Expect top projects to be conference submissions.

Project Ideas

Projects we work on,

- Morphology in language modeling
- In-Document Coreference
- Surface ordering of words in a sentence.
- Question-Answering in Text

Project Ideas

Projects we work on ...

- Morphology in language modeling
- In-Document Coreference
- Surface ordering of words in a sentence.
- Question-Answering in Text

Project Ideas

Projects to consider ...

- Information Extraction from Documents
- Twitter and Social Network Modeling
- Visualization of NLP networks
- Deep-Reinforcement Learning and Languages