Interesting Presentations I met at EMNLP 2018

Zuzana Nevěřilová

November 9, 2018

Outline I

Outline II

- Bernhard Kratzwald et al.: Adaptive Document Retrieval for Deep Question Answering
- Minjoon Seo et al.: Phrase-Indexed Question Answering: A New Challenge for Scalable Document Comprehension.
- Saku Sugawara et al.: What Makes Reading Comprehension Questions Easier?
- Haitian Sun et al.: Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text

Truth or Lie? Spoken indicators of deception in speech I

Julia Hirschberg

Deceptive speech

- deliberate choice to mislead
  - Not considered self-deception, delusion, pathological behavior (pathological liars are impossible to detect), theater, ignorance/error
  - Not considered everyday white lies: very hard to detect (e.g. I loved your talk, I love your new haircut)
- serious lies are easier to detect: because our cognitive load is increased:
  - keep to story straight
  - must remember what we have and have not said
  - fear of detection is increased (we believe that our target is difficult to fool)

Truth or Lie? Spoken indicators of deception in speech II

Julia Hirschberg

Humans are poor at recognizing deception cues(40–60% accuracy)

Current approaches

- automatic methods (polygraph) no better than chance
- human training (behavioral analysis)
- laboratory studies (production and perception, facial expression, body posture, statement analysis, brain activity...)

Objective experiments on human subject to identify spoken language cues to deception

- acoustic-prosodic
- lexical cues
- individual differences (gender, ethnicity, culture...)

Truth or Lie? Spoken indicators of deception in speech III

Julia Hirschberg

Corpus collection

- Columbia SRI Colorado Deception Corpus (2003–)
  7 h of speech, key findings:
  - perform better on male
  - human judges that were high in certain personality features (openness to experience, agreeableness) performed much better
- Columbia Cross Cultural Deception Corpus
  - gender and personality information, compare subjects with different cultural and language backgroundpair people from SAE and China

Experiment

- Setting: questionnaire with true and false answers, recording the subject when they were talking about ordinary things (how they like NY, what restaurants do they go...), lying game, survey

Truth or Lie? Spoken indicators of deception in speech IV

Julia Hirschberg

- Monetary motivation: $1 for interviewer/interviewee
- 340 subjects
- 122 hours of speech
- crowdsourced transcription
- automatic speech alignment
- speech segmented into:
  - inter pausal units
  - speaker turns
  - Q/A sequence
- Each subject undertake the Big5 NEO-FFI personality scores
  - openness to experience
  - conscientiousness
  - extraversion
  - neuroticism
  - agreeableness

Truth or Lie? Spoken indicators of deception in speech V

Julia Hirschberg

- Deception annotation
  - local (keypresses)
  - global deception (linked to the questionnaire)

Machine Learning to Recognize Deception

- Extracted Features
  - text-based: n-gram, linguistics inquiry and word vound (LIWC), word embeddings
  - speech based: openSMILE IS09
  - gender, native language, NEO-FFIs personality scores
  - syntactic features (complexity)
- Classifiers
  - random forest
  - SVM
  - deep learning (BLSTM-lexical + DNN-openSMILE)

Truth or Lie? Spoken indicators of deception in speech VI

Julia Hirschberg

- hybrid approach has the best results: F1 0.64

Improving Deception Detection with Personality Features

- classifier on speaker turns (not IPUs)
- adding personality through multi task learning
  $\rightarrow$ improved F1 from .68 to .744

What can we learn from gender and native language?

- compare distribution of features (interviewees tell lie and are trusted, interviewees tell truth and are trusted, they tell lie and are not trusted, they tell lie and are trusted)
- paired t-test

Truth or Lie? Spoken indicators of deception in speech VII

Julia Hirschberg

Characteristics of deceptive speech

- For all groups: high max pitch, high max loudness – when telling lie
- Some surprising specifics for groups:
  - female have a voice jitter when telling truth
  - Chinese native speakers have higher speaking rate when telling truth
  - Chinese native speakers have higher max pitch when telling lie

In this specific task, ML models perform much better than humans :-o

Truth or Lie? Spoken indicators of deception in speech VIII

Julia Hirschberg

Where do machines better?

- certain speakers?
- groups of languages?
- question types?

Naive Bayes Classifier with 5000 features (syntactic and lexical) to identify what type of lies/liars ML is able to detect
Key findings:

- no correlation: which speakers are easy or difficult to judge
- classifier much better at judging speaker with low conscientiousness
- Easy and hard questions:
  - easiest question(s): have your parents divorced? (only 20
  - the hardest question: have you ever stayed overnight in the hospital as a patient?

Truth or Lie? Spoken indicators of deception in speech IX

Julia Hirschberg

- sensitive questions: who do you love more, your mother or your father?

In case of sensitive questions, humans outperform machines in recognizing deception

Future Work

- create trusted and mistrusted synthetic voices
- Game with a purpose (GWAP): LieCatcher

The Moment of Meaning I

Johan Bos

History and future of computational semantics

- 1970 Montague semantics
- 1980 parsers for small fragements
- 1990 under specification, automated inference
- 2000 wide coverage semantic parsers
- 2010 RTE, large annotated corpora
- 2020 some stuff with neural networks

Why?

- future language technology needs semantic interpretation $\rightarrow$ "explainable NLP"
- improve MT (contradiction checking)
- it's fun!

The Moment of Meaning II

Johan Bos

Example (Lost in Translation)

Please do not empty your dog here!
Nothing sucks like an Electrolux

← something is wrong with the semantics

currently, MT improved a lot but. . . what if the small error (BLEU score still over .9) is a missing negation?
some imprecise translations are OK (e.g. translate a glass of beer as a pint of beer)

Meaning Banking
Parallel Meaning Bank:

- lexical semantics

The Moment of Meaning III

Johan Bos

- formal semantics
- gold-standard meanings
- multi-lingual
- resource for parsing/translation

The resource

- machine produced, human corrected
- language neutral annotation
- use parallel corpora
- English first, de, nl, it
- WordNet/VerbNet/DRT (discourse representation theory)
- bronze, silver, gold data
- accessible at `http://pmb.let.rug.nl`

The Moment of Meaning IV

Johan Bos

boxes: entities, events, time
pipeline: segmentation → parsing → semantic tagging → boxing

Language Neutral Linguistc Analysis

- 1 tagset for each subtask, 1 tokeniser, 1 parser, 1 semantic tagger, 1 boxer
- syntactic component: CCG (combinatory categorial grammar)
  - schemata are grammar rules
- semantic tags: 72 sem tags divided into 13 classes
  - designed in data driven fashion
  - POS tagging is not enough
  - includes NER
  - semantically motivated
  - see: Abdou, EMNLP 2018: What can we learn from semantic tagging?

The Moment of Meaning V

Johan Bos

- compositional semantics (lambda-DRT)
- translation: translate the semantic tags, if needed move the slashes in the syntactic tagging (e.g. if the word order is different)
- copy, merge and split
  - also apply CCG rules
- example: I do like ice cream. Ich mag wirklich Eiscreme.

Boxing Day

- Discourse representation structure
- very similar to AMR (LISP -> Prolog the right way... ;-)
- DRS differ from AMR
  - scopes, recursive structures
  - see Discourse Semantics with Information Structure

The Moment of Meaning VI

Johan Bos

- Tom is stuck in his sleeping bag.
  Tom: male, sleeping-bag: object
- List of Phenomena (PMB): 42 items (because 42 is the answer for everything :-) but the number is growing

Drowning by Numbers

- Evaluating Meaning Representations
  - check for logical equivalence
  - provers for FOPL
  - discrete score (0=no proof, 1=proof)
- syntactic evaluation
  - check matching tuples
  - continuous score
- DRS: clause notation (e.g. 8 out of 9 clauses match)

The Moment of Meaning VII

Johan Bos

The match

- Classic Boxer vs. Neural Boxer
- Neural Boxer: no tokenisation, OpenNMT, 2BiLSTM, 300 nodes, naive dropout, general attention, beam size 10 during decoding
- How does the neural boxer learn?
  - variables as nameless dummies
  - remove variables, replace second mentions to relations to variable introduced
  - character based input + output (s,h,e,+,s,h,o,w,e,r,s,+,e,v,e,r,y,+,m,o,r,n,i,n,g)
- Neural boxer has F1 78, classical has 74, neural boxer with silver data has 84

The Moment of Meaning VIII

Johan Bos

The Silence of the Lambdas (dissapointment after Neural Boxer beats the Classical one)

- seq2seq model, remove spaces
  (s,h,e,s,h,o,w,e,r,s,e,v,e,r,y,m,o,r,n,i,n,g)
  $\rightarrow$ decrease 5% in F-score

The Moment of Meaning IX

Johan Bos

Future

- Meaning Banking
- Explainable NLP: E.g. if a RTE system finds a contradiction, is it able to explain why there is a contradiction?
- Is the meaning representation a method how to explain meaning?
- MTL with semantic tagging as aux task?
- DRS parsing in a nutshell: join us in the shared task, Gothenburg
- `pmb.let.rug.nl`

Data For Clinical Text Mining (invited talk) I

Goran Nenadic

- data availability: major issue in clinical text mining
- publicly available datasets: scientific papers, drug reviews, discharge summaries etc.
- Thyme, Bioasq, MIMIC, PubMed, CRIS, SNPPhenA, EU-ADR, SemEval 2013 DDI (DrugBank, MEDLINE), ADE-EXT, reACE
- many people use local datasets

See also Citizens' jury in June 2018: What do patients say?

Data availability

- automated large scale de-identification

Data For Clinical Text Mining (invited talk) II

Goran Nenadic

- worst precision/recall is in de-identifying organizations and professions
  - large variety in occupation types and expressions that can be used
  - entity recognition is context-dependent: Edwards can be the doctor, patient, or Carpentier-Edwards Aortic Valve
- what happens to entities during de-identification
  - keep (e.g. doctor's name or disease name)
  - redact (e.g. change patient's name)
  - map (remove some information, e.g. instead of a precise date: November Weekday 40-50 years)
- role specific processing: e.g. not to remove doctors names
- rare item set analysis: check the data (mostly ad hoc)
- Sampling is important: not necessarily to process the whole data

Data For Clinical Text Mining (invited talk) III

Goran Nenadic

Data availability: solutions

- Synthetic data:
  - complement existing datasets
  - CLEF eHealth Evaluation Labs 2015/16
  - quality of generated data
    - are they clinically correct?
    - is privacy preserved?
    - some other metrics?
- Veterinary data (savsnet.co.uk)
  - but animals vomit a lot (or better if they do, people take them to the doctor)
  - learning related terms
  - also contain sensitive data (about the owner)

Data For Clinical Text Mining (invited talk) IV

Goran Nenadic

Supervised vs. unsupervised (see a paper from 1968: Fashion in Science: does it exist?)

State-of-the-art in healthcare text mining (clinical NLP)

- sharing methods and models: reproduce and reuse (so far it is not very usual)
- FAIR = findable, accessible, interoperable, reusable
- report a minimum set of information (guidelines, IAA, approach...)
- technical aspects: GATE, executables, Jupyter notebooks, repo?
- cultural aspects: collaborative science?
- ethical aspects: de-identification issues, applying neural networks on encrypted data

Data For Clinical Text Mining (invited talk) V

Goran Nenadic

- user-centric system
- transparency: interpreting the results of NLP systems (blackbox)
- gentle NLP (not to tag, extract, predict ... if the system is not confident enough)

- modelling normal state
  "2 hours last night – that's a new record" = 2 hours are not normal, it means insomnia

Community: `http://healtex.org`

Trying to Understand RNN for NLP (invited talk) I

Yoav Goldberg

the most important word is understand = the gap between ML and NLP

History

- in 1950-1990 we were writing rules (transparent)
- 1990-2000 we were using corpora
- 2000-2014 machine learning
- 2014- neural networks (complete blackbox)
- 2021+ writing many rules (back to the roots) aided by ML :-)

Trying to Understand RNN for NLP (invited talk) II

Yoav Goldberg

Current approach: No matter what the task, you throw a BiLSTM at it and it will do the stuff...
What to do with LSTM: use them, build tools for them, understand them (even though reviewers don't care much)

Q1 what is encoded/captured in a vector?
paper: Fine grained analysis of sentence embedding using auxiliary prediction tasks
Q2 what kinds of linguistic structures can be captured by an RNN?
paper: Assessing the ability of LSTMs to learn syntax sensitive dependencies
Q3 how did a given model reach a decision? how is the architecture capturing the phenomena?
paper: Sharp Nearby, Fuzzy Far Away: How Neural Language Models Use Context

Trying to Understand RNN for NLP (invited talk) III

Yoav Goldberg

Q4 when do models fail? what can't they do?

The "build it and break it" contest

paper: Breaking NLI Systems with Sentences that Require Simple Lexical Inferences

Q5 what is the representation power of different architectures?

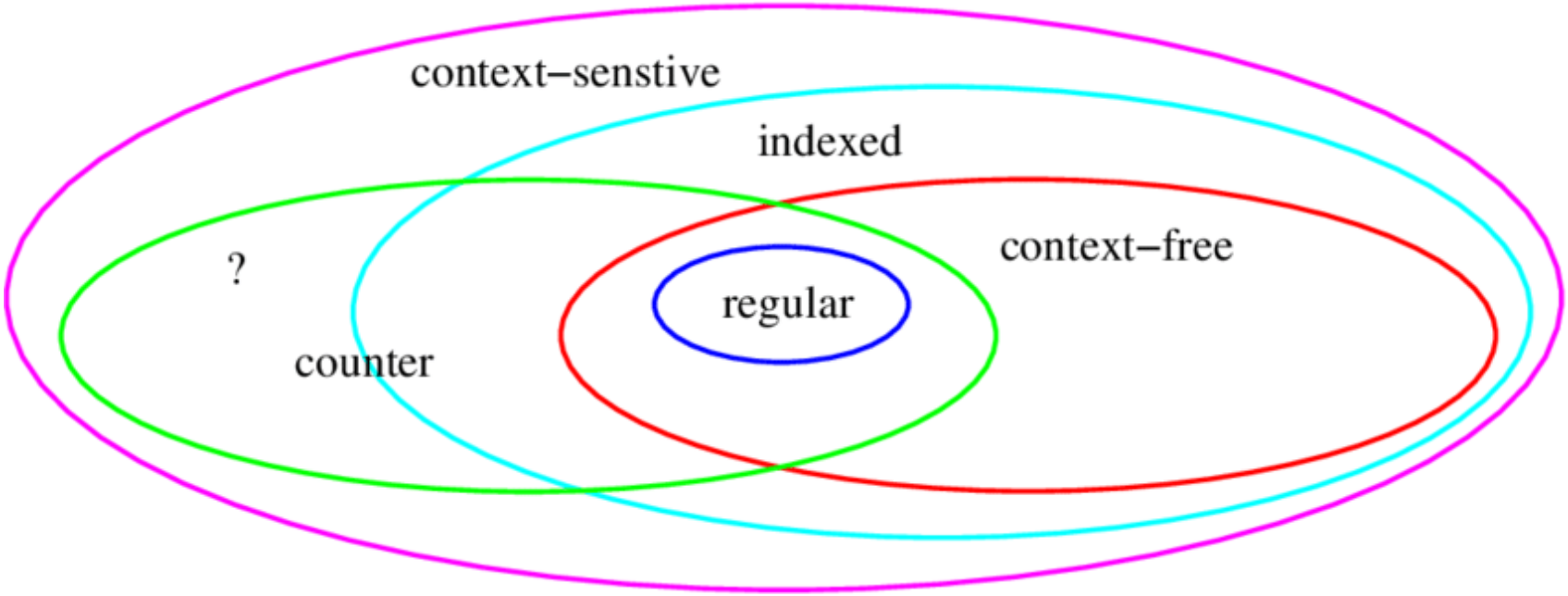
paper: Recurrent Neural Networks as Weighted Language Recognizers Formal expressive power of RNNs

- are all RNN equivalent? do RNNs have Turing power? Yes. The proof requires infinite precision (Not very useful answer though).
- In reality, RNN construction requires extra processing time at the end of the sequence (which makes them less powerful)

Trying to Understand RNN for NLP (invited talk) IV

Yoav Goldberg

- different RNN flavors: Elman RNN (SRNN) with saturating activation, IRNN with ReLU activation, Gated RNNs – do they all have the same expressive power? No.



With finite precision (real time) Elman RNN are finite state.
Gated (GRU, LSTM) are better than non-gated (SRNN, IRNN).

Trying to Understand RNN for NLP (invited talk) V

Yoav Goldberg

In Chomsky hierarchy, counters are stronger than regular languages but less strong than context free (they can do $a^n b^n$ but they can't do palindromes)

For more information, see 1968 paper: Counter machines and counter languages

- GRU/SRNN $\approx$ regular grammars
- LSTM/IRNN $\approx$ counters
- Counting is easy, it's just to saturate 3 gates. GRU/SRNN cannot count.
- Tranformers are less powerful than LSTMs.

Trying to Understand RNN for NLP (invited talk) VI

Yoav Goldberg

LSTM vs. GRU

- $a^n b^n$ with LSTM and GRU respectively: LSTM learned the concept, GRU did not
- $a^n b^n c^n$ similar results

Small architectural choices can change the expressive power.

Are LSTMs needed for languages? Are GRUs not enough?
Q6 Extracting a discrete representation from a trained model
Extracting (finite state automata) FSAs from RNNs . . . unfinished

How to understand language (BlackBoxNLP) I

Leila Wehbe

understanding of neural networks has already happened in vision

Experiments on brain and language

- traditionally, contrast based experiments
  - condition 1 (abstract words), condition 2 (concrete words), brain area response
  - condition 1 (semantically surprising), condition 2 (semantically unsurprising)
  - ...
  - expensive, infinite binary contrasts?

How to understand language (BlackBoxNLP) II

Leila Wehbe

Where NLP representations can help model brain representations

- observe brain activity when subjects do ordinary activities
- subjects sits in front of a scanner, is asked to read something and the brain activity is scanned (somehow naturalistic)
- different words activate different brain area
- short contexts (one word, previous word) vs. long contexts (sentences and more)
- build a predictive model for brain activity
- How does the brain combine multiple words?
- Does RNN activation correspond to brain activity?
- Are we able to detect when and where some processes in the brain occur (visual, characters, semantics, syntax, motion. . .)?

How to understand language (BlackBoxNLP) III

Leila Wehbe

Word and sentence embeddings

- compare embeddings with the brain data
- what does the alignment tell us?

Model a blackbox (brain) with a blackbox (NN) – does it make sense? what is important in brains?

- spatial structure is important, is not random
- time is important

The experiment with Magnetoencephalography (MEG)

- $\approx$ 100 sensors of the MEG can read signal every one millisecond $\rightarrow$ high temporal resolutions
- measuring MEG activity as a function of stimulus text
- predict the activity based on word embeddings
- consider differences between different subjects

How to understand language (BlackBoxNLP) IV

Leila Wehbe

Observations

- Layer representations are correlated with each other
- embedding vector: word length + POS + maybe other features
- ELMo: some embeddings are better at predicting MEG activity than others
- NN embeddings closely track MEG activity
- shallow, no good for looking at long range (sentences and more)
- fMRI is better than MEG at longer time scales
- ELMo seems particularly well suited for fMRI, actually ELMo has both long and short range

Note: ELMo = `http://allennlp.org/elmo`

Reasoning about Actions and State Changes by Injecting Commonsense Knowledge I

Niket Tandon, Bhavana Dalvi, Joel Grus, Wen-tau Yih, Antoine Bosselut and Peter Clark

Procedural text

- example: photosynthesis desription (entities: water, light, co2, sugar)
- inferences:
  - roots absorb water from the soil → water is at the roots
  - the water flows to the leaf → water is at the leaf
- infer not only the state but also the changes (the process)
- what to do if some of the entities are missing?

Dataset

- how can we teach computers? corpus
- ProPara Dataset: how to use dishwasher? how do volcanoes work? ...

Reasoning about Actions and State Changes by Injecting Commonsense Knowledge II

Niket Tandon, Bhavana Dalvi, Joel Grus, Wen-tau Yih, Antoine Bosselut and Peter Clark

- multiple paragraphs on topic with entities of interest
- paragraph annotated with actions and state changes

Models

- simple neural model
- input – output: paragraph + entities $\rightarrow$ entity states
- paragraph + enitites (encoder) $\rightarrow$ action encodings $\rightarrow$ (greedy decoder) $\rightarrow$ state changes predictions

Encoder

- action encodings + bilinear attention
- greedy decoding may result in nonsensical predictions
- example natural constraints:
  - entity must exist before it can be moved

Reasoning about Actions and State Changes by Injecting Commonsense Knowledge III

Niket Tandon, Bhavana Dalvi, Joel Grus, Wen-tau Yih, Antoine Bosselut and Peter Clark

- roots typically do not move
- Commonsense about actions, state changes
  - global hard constraints
  - Topic + SRL + action rules (VerbNet) → commonsense for Topic
    (SRL = semantic role labeling, VerbNet = verb valency dictionary)
  - ProStruct Model: treat all possible options, remove those that violate the global constraints, find a goldpath

Evaluation

- 4 questions on the ProPara dataset
- rulebased: 35,9 F1

Reasoning about Actions and State Changes by Injecting Commonsense Knowledge IV

Niket Tandon, Bhavana Dalvi, Joel Grus, Wen-tau Yih, Antoine Bosselut and Peter Clark

- after adding commonsense significantly improves precision, slightly drops recall

Error analysis

- implicit reference
- coreference resolution
- knowledge retrieval
- The dataset is public: `http://github.com/allenai/propara`

Paper: `http://aclweb.org/anthology/D18-1006`

Textual Analogy Parsing: What's Shared and What's Compared among Analogous Facts I

Matthew Lamm, Arun Chaganty, Christopher D. Manning, Dan Jurafsky and Percy Liang

Motivation: Visualizing Quantitative Text instead of search for facts

Example: According to the us census today only 10 % White Americans live at or below the poverty line

- source: US Census
- quantity: live at or below the poverty line
- time: today
- value: 10
- whole: White Americans

Graph Based Model

- fact $\rightarrow$ analogy

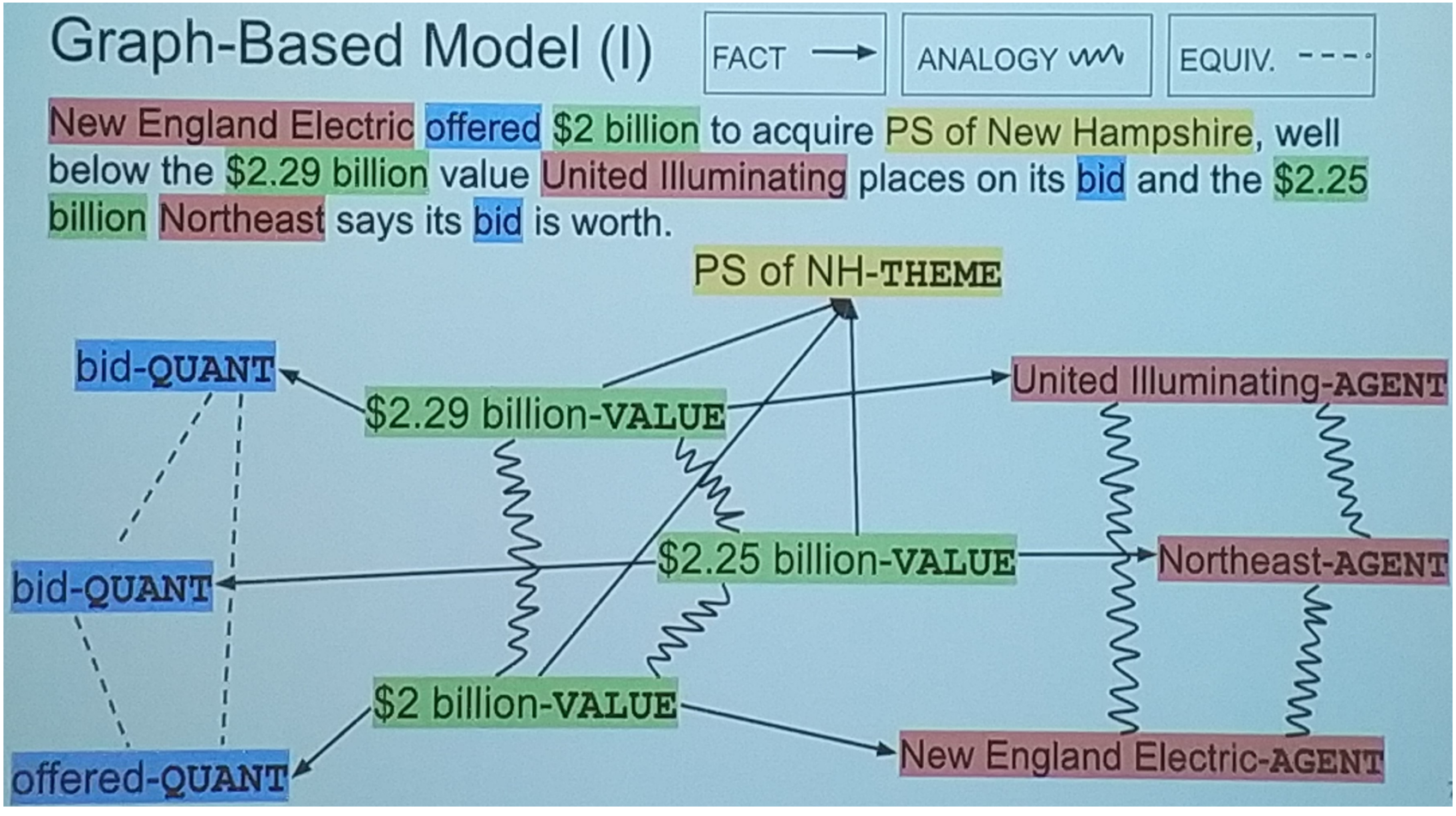
Textual Analogy Parsing: What's Shared and What's Compared among Analogous Facts II

Matthew Lamm, Arun Chaganty, Christopher D. Manning, Dan Jurafsky and Percy Liang

- TAP has to resolve two challenging types of relations:
  - shared content: scope/gapping = single syntactic element serves as a role in multiple QSRL frames
  - compared content: synonymy/coref = multiple elements appear in a sentence but contribute the same role to the shared content (e.g. bid-offer)

Textual Analogy Parsing: What's Shared and What's Compared among Analogous Facts III

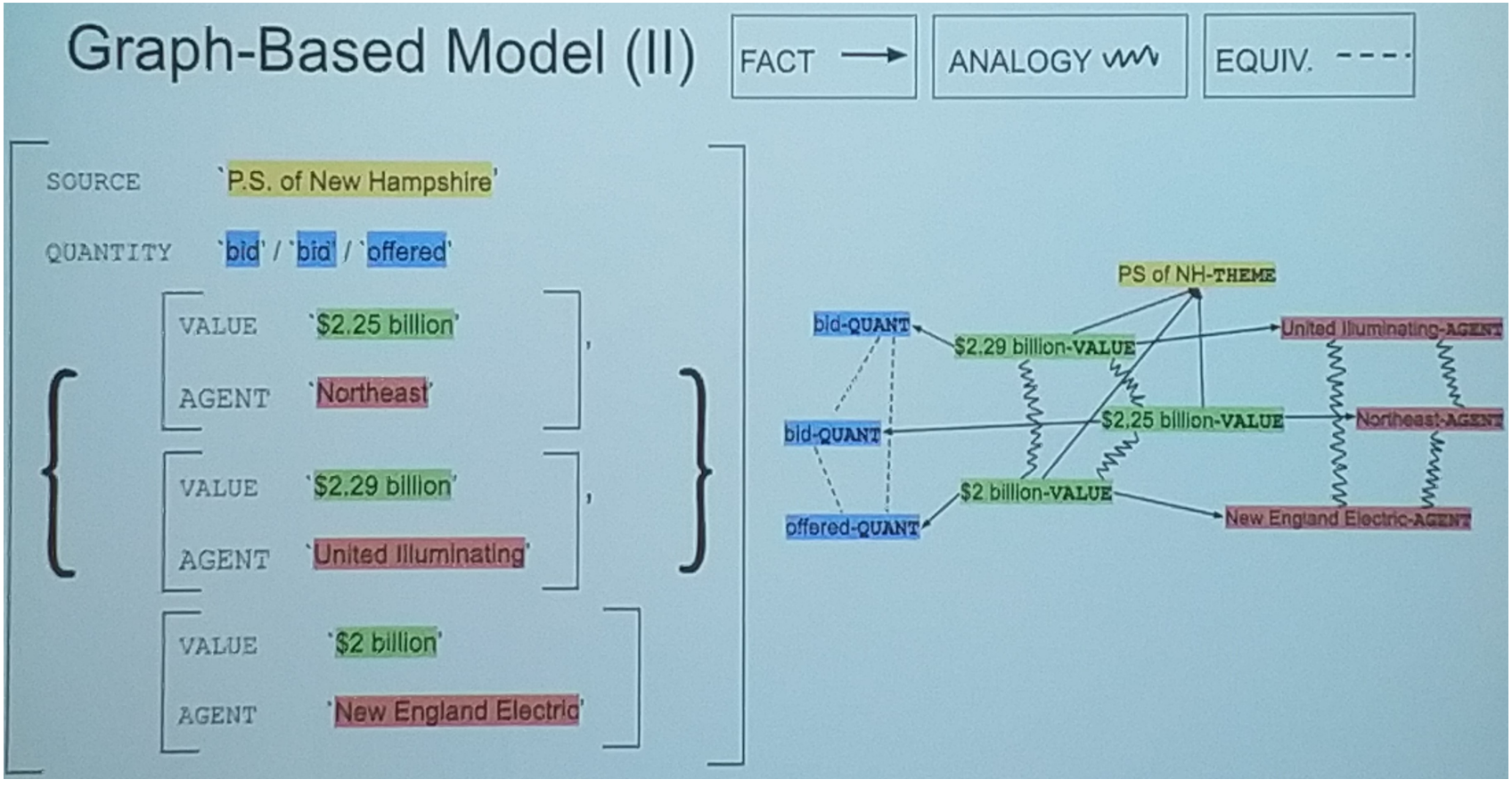
Matthew Lamm, Arun Chaganty, Christopher D. Manning, Dan Jurafsky and Percy Liang



Textual Analogy Parsing: What's Shared and What's Compared among Analogous Facts IV

Matthew Lamm, Arun Chaganty, Christopher D. Manning, Dan Jurafsky and Percy Liang

From analogy graph to analogy frame



Textual Analogy Parsing: What's Shared and What's Compared among Analogous Facts V

Matthew Lamm, Arun Chaganty, Christopher D. Manning, Dan Jurafsky and Percy Liang

Neural + ILP Architecture (integer linear programming)

- span prediction (CRF – conditional random fields), edge prediction, decoding
  - token prediction → span prediction
  - edge prediction between two spans (Roth, Lapata 2016)
  - decoding into analogy frames using ILPs (analogy can only be between similarly labeled entities)
- post ILP can handle many errors
- `http://github.com/mrlamm/textual-analogy-parsing`

Paper: `http://aclweb.org/anthology/D18-1008`

A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification I

Mounica Maddela and Wei Xu

Simplification tasks mostly based on corpus information: frequency and word length.
Weakness: "more frequent words are simple" is a wrong assumption, e.g. "foolishness" is simpler than "folly"

A large word-complexity lexicon

- 15K most frequent words rated on a 6 point Likert scale (very simple, simple, moderately simple, moderately complex, complex, very complex)
- 11 non native annotators (graduate students, Chinese, Russian, Indian ...)
- 2.5 hours to rate 1000 words
- 41% simple, 30% intermediate, 19% very simple

A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification II

Mounica Maddela and Wei Xu

- inter-annotator agreement (IAA): 0.64 (Pearson κ) = relatively high agreement
- difference on vs. rest: <0.5 for 47%, <1.0 for 78%, <1.5 for 93% annotations

Pairwise Neural Ranking Model

- input word/phrase pair (e.g. "adversary" – "enemy")
- feature extraction (no. of syllables, word length...), Gaussian-based feature vectorization (10D vectors), multilayer perceptron
- Evaluation on English Lexical Simplification Shared Task (SemEval2012)

A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification III

Mounica Maddela and Wei Xu

- Gaussian based feature vectorization improves the results significantly
- Also improvement in paraphrase generation

Polysemy

- The target words were identified within 10 sentences
- tha annotation is without any context

Paper: `http://aclweb.org/anthology/D18-1410`

Adaptive Document Retrieval for Deep Question Answering I

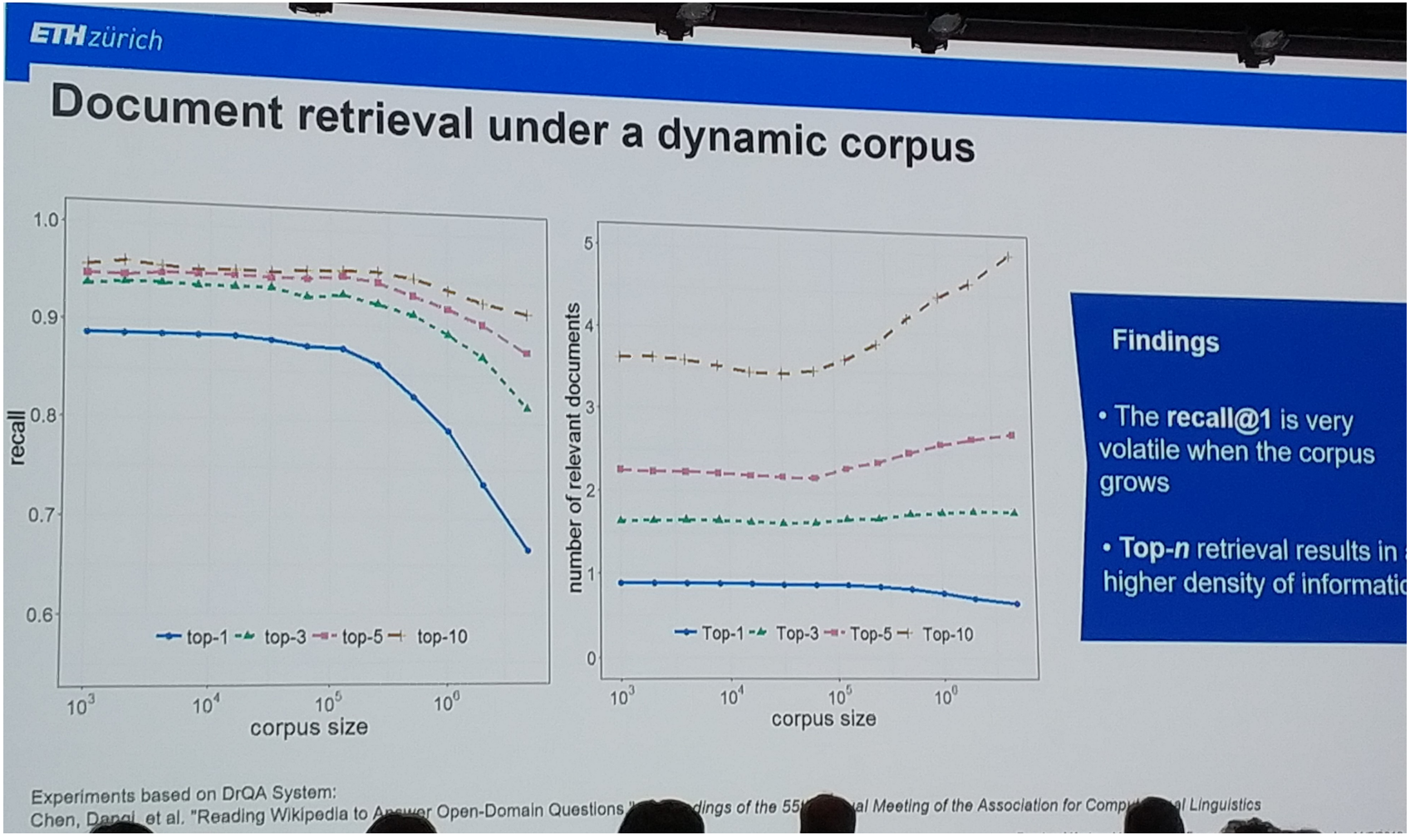
Bernhard Kratzwald and Stefan Feuerriegel

content based QA systems with neural answer extraction:

- for small corpora, top-1 system outperforms any other configuration
- things change at corpus size $> 10^6$

Adaptive Document Retrieval for Deep Question Answering II

Bernhard Kratzwald and Stefan Feuerriegel



Adaptive Document Retrieval for Deep Question Answering III

Bernhard Kratzwald and Stefan Feuerriegel

- selecting more paragraphs $\rightarrow$ higher performance
- select $>\approx$70 paragraphs $\rightarrow$ lower performance
- Threshold baseline: the mode we're confident, the less documents we select
- Learn the cut-off point
- The adaptive approach is numerically more stable and robust than selecting a particular number.
- `http://github.com/bernhard2201/adaptive-ir-for-qa`

Paper: `http://aclweb.org/anthology/D18-1055`

Phrase-Indexed Question Answering: A New Challenge for Scalable Document Comprehension. I

Minjoon Seo, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi and Hannaneh Hajishirzi

- Q + A get to some domain: extractive datasets
- but what about open domain QA?
  - Wiki space too large
  - reduced by Information retrieval
  - IR data go to the model
  - wrong document from IR propagates the error to wrong answer

Index phrases

- document indexing + nearest neighbor (NN) search
- given a document, we extract phrases, phrase encoding, question encoding, NN search

Phrase-Indexed Question Answering: A New Challenge for Scalable Document Comprehension. II

Minjoon Seo, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi and Hannaneh Hajishirzi

$\hat{a} = argmax_a F_\theta(a, q, d)$ $\leftarrow$ decomposability is a strong constraint

phrase indexed QA (PIQA) phrase embedding `http://pi-qa.com`
Evaluation: BERT 92% F1, SA Elmo 86%, SA+ Elmo 64%,

decomposability gap between the two ELMos

Paper: `http://aclweb.org/anthology/D18-1052`

What Makes Reading Comprehension Questions Easier? I

Saku Sugawara, Kentaro Inui, Satoshi Sekine and Akiko Aizawa

Datasets: MCTest, bAbI, SQuAD, TriviaQA, ARC, Clicr and many other datasets

For reading comprehension: many skills (coreference, ...) are needed.
But do MRC datasets really require comprehension?

NLU tasks contain unintended patterns

- word/patterns features specific to certain answer classes
- adversarial resources try to fool the answers
  for example, if there is only one answer candidate, then it's easy to find the correct answer
- two heuristics to identify easy and hard questions
  - how many q are solved with only first k tokens?

What Makes Reading Comprehension Questions Easier? II

Saku Sugawara, Kentaro Inui, Satoshi Sekine and Akiko Aizawa

- how many q have theirs answers in the most similar sentence?
  - QAnga and multiple choice show small difference between full and k=1
  - answer in the most similar sentence is strong in SQuAD, AddSent, NewsQA...
- hard q = k>=2 and answer is not in the most similar sentence

Many datasets are unbalanced (e.g. in QAnga, easy qs are easy, hard qs are hard and rare)

Paper: `http://aclweb.org/anthology/D18-1453`

Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text I

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov and William Cohen

structured and unstructured knowledge

- DrQA: read Wiki $\rightarrow$ retrieve text $\rightarrow$ read text
  high recall, difficult to extract relevant part of the text
- QA with structured knowledge base (KB)
  semantic parsing , e.g. Neural Semantic Parsing (Liang et al, ACL 2107)
  - late fusion (train models separately)
  - early fusion (combine all knowledge, train models)

Open Domain Question Answering Using Early Fusion of Knowledge Bases and Text II

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov and William Cohen

GraftNet: represent text and KB (early fusion)

- instead of direct edges between Meg Griffin – Lacey Chabert, they added nodes such as `voiced-by`
- embedding propagation using Freebase and Wikipedia
  - self
  - KB
  - text

Fact Dropout

Is early fusion better than late fusion?

If KB coverage is high, KB is better than text. In this case late fusion has worse performance because of noise.

`http://github.com/OceanskySun/GraftNet`

Paper: `http://aclweb.org/anthology/D18-1455`