Notes on NAACL 2018

Zining Zhu *

New Orleans, Louisiana, June 5, 2018

Contents

*Winterlight Labs + University of Toronto, http://ziningzhu.me/2018/06/05/NAACL2018, zining.zhu@mail.utoronto.ca

1 Friday 0601

1.1 T1: Modelling NL, programs, and their intersection

Speakers Professors Graham Neubig and Miltos Allamanis for sharing the slides.

1.1.1 Programming language vs natural language

- a lot of similarities between programming languages and domain-specific languages
- data sources available. e.g: Stack Overflow
- Categories of data: intent, written intent, code snippets, doc strings, comments, diff messages

1.1.2 Methods for mapping code to natural language

- Translation: (natural language description). Methods: machine translation, CNN + attentions, etc.
- Code summarization. e.g: Iyer et al "summarizing source code using a neural attention model". e.g: predict method names
- Convolutional neural attention models with attention mechanisms (which decide whether copy or summarize, similar to pointer-generator network)
- Incorporating the execution results to evaluate quality of generated programs
- Programming by demonstration...?
- Semantic parsing from Q-A pairs. Weak supervision is easier to create (e.g: for generating SQL. Zhong+17, Clarke+10)

1.1.3 Program generation: map from language to code

- Machine translation, but with clear destination syntax rules
- Historical methods: rule-based transformations, grammar-based models, neural models. Following talks about neural approaches.
- How to take advantages of features of code? e.g: copy variable names. (word-level or character-level, or tree-level (Dong+16)) e.g: Top-down generation of CFG rules
- Also possible to generate codes from coarse-to-fine level (Dong+18). First predict the sketches, and then codes
- Code synthesis with natural language guidance (Polosukhin+18)
- Reconstruction loss: supervision without execution (Yin+18). Can use VAE formulation
- Code search: output API calls, Gu+2016

1.1.4 Modeling natural language aspects of source code

- Predict variable names
- Type inference. This has a lot to do with intelligent IDE
- If we represent program structures as a graph.

1.1.5 Modeling communicative aspects of software projects

- Model discussion topics: what are they talking about?
- Measuring the complexity of languages / codes (meaningful to Q-A sites)
- sentiment analysis for software (Lin+18)

1.2 T3: Scalable construction adn Reasoning of Massive KB

Speakers Professors Xiang Ren, Nanyun Peng, and William Yang Wang

Intro

1.2.1 Use cases for Text to Structure

TripAdvisor travel review, precision medicine (read the PubMed papers), search engines.

Prior art: extracting structures with repeated human effort. Works pretty well but hard to scale.

Our method: effort-light structure extraction. Knowledge -> text corpus -> corpus-specific models -> structures.

Difficulty: aparsity of "matchable" (incomplete knowledge bases, low-confidence matching),

1.2.2 Methodology

The tasks

- Data-driven text segmentation
- Learning corpus-specific model
- Structures from the unlabeled data

1.2.3 Part 1: Recognize entities of target types in text

Traditional NER systems: sequence model training. e.g: Stanford NER, Illinois name tagger, IBM Alchemy APIs

Training sequence models is slow + heavy reliance on corpus-specific human labeling.

Weak-supervision systems: pattern-based bootstrapping (send several examples as "seeds"). Problem: can include wrong patterns.

Leveraging distance supervision. 1. Detect entity names from text
2. Match name strings to KB entities.
3. Propagate types to the un-matchable names

LImitations:

- Context-agnostic type prediction.
- Sparsity of contextual bridges (people describing the same things using different terms). This results in inefficient type propagation.

Example: ClusType approach (KDD '15): type propagation + relation phrase clustering at the same time.

Smoothness assumption: if two nodes are similar saccording to the graph, then their type labels should also be similar.

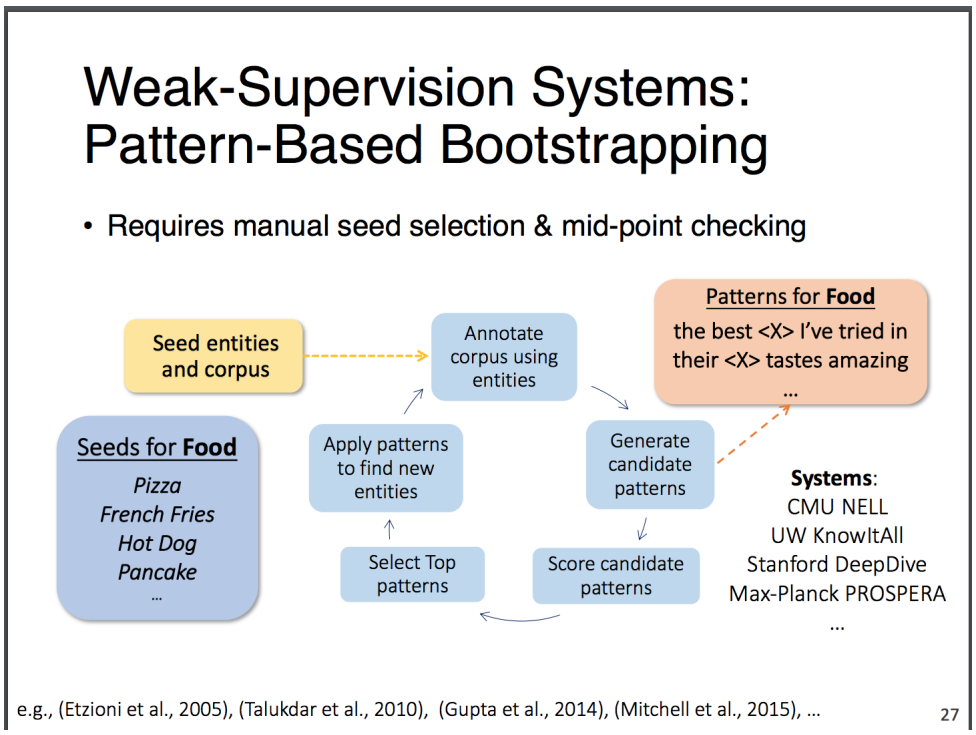
Two relation phrases should be grouped together if: (1) similar string; (2) similar context; (3) similar types for entity arguments. –> Multi-view clustering.

From coarse-grained typing to fine-grained entity typing: For a clean mention, its "positive types" should be ranked higher than all its "negative types".

Hierarchical type inference (Ren et al EMNLP '16)

Partial label embedding (PLE, KDD '16)

Comparison: WSABIE (Google ACL '15) Predictive Text Embedding (MSR)



Prior works e.g: CoType approach (WWW '17) Co-embedding for typing entities and relations

1.2.4 Part 2: Joint extraction of typed entities and relations

How to leverage other knowledge, such as the distributional statistics of characters and words, and annotations for other tasks and other domains, and the linguistics and problem structures, to combat the problem of inadequate supervision and conduct low-resource information extraction.

Traditional NER method sequence tagging models, hand-engineering features

Neural NER models e.g: RNN for representation.

Distributional similarity of words Why not perform joint learning of word embeddings and NER?

e.g: chinese word boundaries

Sharing high-level representations (Peng and Dredze, 2016)

Domains for languages Multi-task multi-domain learning. (Peng and Dredze, 2017)
Task-specific models – domain projections – shared representation learner

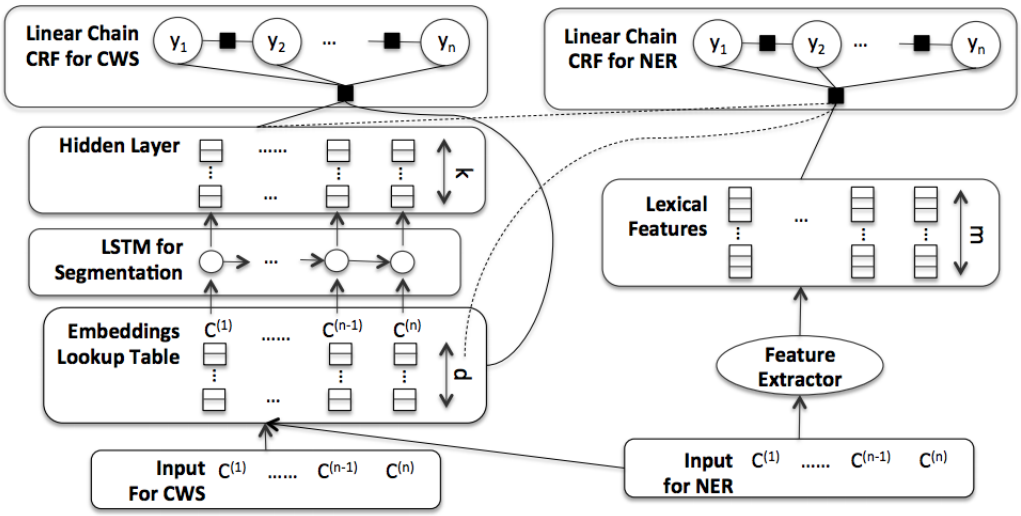


Figure 1: The joint model for Chinese word segmentation and NER. The left hand side is an LSTM module for word segmentation, and the right hand side is a traditional feature-based CRF model for NER. Note that the linear chain CRF for NER has both access to the feature extractor specifically for NER and the representations produced by the LSTM module for word segmentation. The CRF in this version is a log-bilinear CRF, where it treats the embeddings and hidden vectors inputs as variables and modifies them according to the objective function. As a result, it enables propagating the gradients back into the LSTM to adjust the parameters. Therefore, the word segmentation and NER training share all the parameters of the LSTM module. This facilitates the joint training.

How to build NER for a new language using (1) comparable corpora (e.g. wikipedia) and (2) English NER tagger? (Want, Peng and Duh, 2017)
Motivation: learn a bilingual word embedding

Two approaches:

- Fixed embeddings
- multi-task training

Encoding linguistic structures to improve e.g: cross-language N-ary relation extractions. Problem: hard to define the shortest path. Also, we are not allowed to go across the boundaries. (Peng et al, 2017) representation learning framework.

• Goal; want to construct a representation learner, that captures difference types of dependencies over an *acyclic graph*.

• Previous approaches: graph neural network, tree neural network, etc. Problem: RNNs are expensive, and that information does not propagate to distant nodes.

• (Peng et al, 2017) cross-sentence n-ary relation extraction with graph LSTMs (in comparison to chain LSTM, there is one more forget gate per dependency)

• Multi-task learning from the shared representation learning

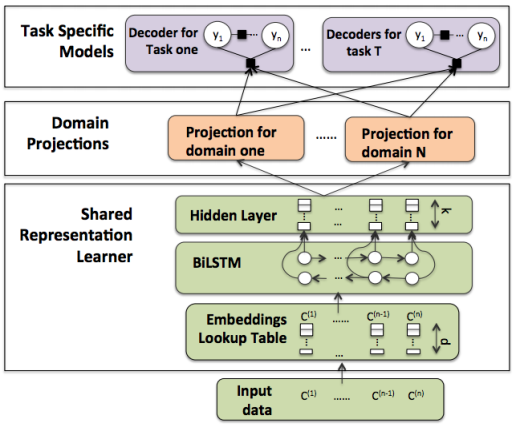


Figure 1: An overview of our proposed model framework. The bottom layer is shared by all tasks and domains. The domain projections contain one projection per domain and the task specific models (top layer) contain one model per task.

1.3 Part 3: Recent advances in knowledge base reasoning

Motivations Knowledge Graphs are not complete. Missing links, etc.

- Knowledge graph supports various applications: structured search, QA, ASR, relation extraction, summarization, etc.
- Goal: complete the knowledge graph automatically (leveraging existing knowledge graph).

Path-based reasoning Why do we need path-based algorithms? (but not neural network embeddings) Explainability!

- Path-ranking algorithm (Lao et al 2011) First random walk with restarts, then do LogReg to rank different paths (make paths leading to the correct destinations have higher weights)
- ProPPR, Wang et al 2013 PhD thesis and Want et al 2015. Generalizes PRA with recursive probabilistic logic programs. May use other relations to jointly infer this target relation.
- Subgraph feature extraction, Gardner et al 2015
- Chains of reasoning. Das et al 2017. PRA to derive path, then use RNNs to perform reasoning of the target relation.

Embedding-based reasoning Related method. (Robust and scalable)

- RESCAL, Nickel et al, 2011. Tensor-based factorization. Head entity - tail entity - relation tensor. $Y = EWE^T$
- TransE, Bordes et al, 2013. If you have the initial embedding, and you add the relation to the head entity, you should get close to the target tail entity.

- Neural Tensor Network, Socher et al, 2013
- TransR / CTransR Lin et al 2015
- Complex Embeddings, Trouillon et al, 2016
- Poincaré embedding. Get out of the Euclidean space. Learn hierarchical KB representations by looking at hyperbolic space.

$$d(u,v) = arcosh(1 + 2\frac{||u-v||^2}{(1-||u||^2)(1-||v||^2)})$$

- ConvE (Detters et al, AAAI 2018) learn entities with CNN. Reshape head and relation embeddings into "images".

Bridging path-based and embedding-based reasoning : DeepPath, MINERVA, and DIVA

- RL for KB reasoning: DeepPath (Xiong et al 2017 EMNLP). Path finding as a MDP. Train RL agent to find paths. Represent KG with pretrained KG embeddings. Use he learned paths as logical formulas.
- MINERVA: Das et al ICLR 2018. Go for a walk and arrive at the answer.
- DIVA: Variational KB reasoning. Inferring latent paths connecting entity nodes.

Sidenote: RL is a general purpose framework for decision making.

1.4 T5: Socially Responsible NLP

Speakers Professors Yulia Tsvetkov, Vinodkumar Prabhakaran, Rob Voigt
(ref: CMU CS11830)

• Be careful: nobody is expert simultaneously in all of the sociology + psychology + linguistic + CSC + ML + statistics.

1.4.1 Ethics in NLP: foundations

• What is ethics? About doing the good / right things. Problem: sometimes cannot define good / bad properly.
• Another example: the chicken classifier (hen -> egg farm; roster -> meat farm)
• Ethics versus law
• Identify a range of problems / questions we should ask when building NLP systems.
• E.g: the A.I. "Gaydar"

1.4.2 Technical Aspects

Humans are the "natural" in NLP In a way, NLP is human subjects research.
• Self-selection bias. e.g: who posts on Yelp
• Reporting bias. e.g: People do not necessarily talk about things in the world in proportion to their empirical distributions.
• (Jurgens et al ACL 17) Socioeconomic bias in language identification.

1.4.3 Case study

• The semantic of words contain inherent biases. e.g: the bias embedding test. Might be able to exclude some of them using fair learning.
/* Excluding gender differences from semantic embeddings is possible, but how about languages like French, where gender difference is encoded as syntactic rules? */

1.4.4 Test

• Should care about whether the task is beneficial to the people involved. The purpose is not going to build "gaydar" or "tell the race of driver based on the police officer's speeches".

• There can be multiple causes for an effect. We should not give blatant judgements without fully assessing them. For example, among those pulled over *only for minor ticketing*, African Americans receive more tickets for minor car damages. This could due to biases in police officers. This could also due to their economic status (less frequently go to repair the cars once damaged?), etc.

• There are complicated reasons behind these problems. Be careful when organizing sentences, etc.

• Extension: CS294 fair learning. Also Graeme Hirst's CSC D03 (social impacts of ML).

Takeaway People are focusing more on the fairness of machine learning. ML researches should take more social responsibilities: Present the researches in an explainable manner, and explore the indications of the researches.

2 Saturday 0602

2.1 Information Extraction 1

Empire A

2.1.1 [19] Joint bootstrapping machines for high confidence relation extraction

- Challenge: semantic drift
- solution: BREX. Use entity and template seeds jointly

2.1.2 [39] Label-aware double transfer learning for cross-specialty medical NER

- Problem: NER from electronic medical records
- Framework: See figure
- Optimization goal: $\mathcal{L}_{]\nabla\{}+\alpha\mathcal{L}_{La-MMD}+\beta\mathcal{L}_{\surd}+\gamma\mathcal{L}_{\text{regularizer}}$, CRF + La-MMD loss + parameter similarity loss + regularization.

2.2 Morning Poster: Discourse and Pragmatics 1

Elite Hall A

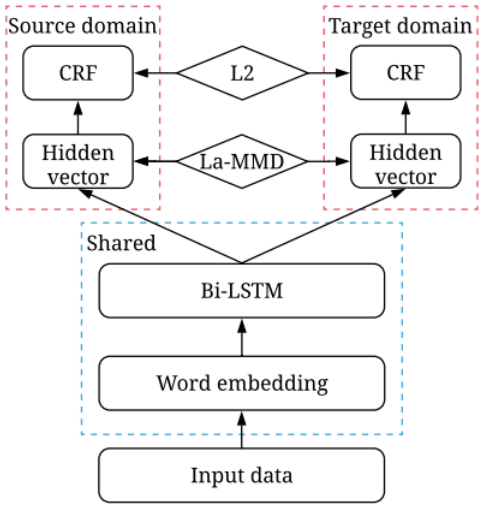


Figure 1: La-DTL framework overview: embedding and Bi-LSTM layers are shared across domains, predictors in red (upper) boxes are task-specific CRFs, with label-aware MMD and L2 constraints to perform feature representation transfer and parameter transfer.

2.2.1 [34] Multi-task learning for argumentation mining in low-resource settings

- Task: argumentation mining: segment a text into argumentative and non-argumentative components and identify them.
- : Method: MTL (training a system to solve several conceptually different AM tasks jointly) improves performance over learning in isolation.

2.2.2 [14] Natural answer generation from heterogeneous memory

- Task: Seq2seq sentence-in sentence-out QA
- Problem: Information come from heterogeneous information sources.
- Proposed model: Incorporate three components in the decoder hidden state: h_n, predicting words from the vocabulary, h_k: key pointer, h_v, value pointer. Use a gate to mix them, so the resultant network is optimizable through back propagation.

2.2.3 [3] Integrating stance detection and fact checking in a unified corpus

- Problem: (part of) fact checking. Decide whether a claim is relevant to a document, and decide whether the document supports the claim.
- This work describes the corpus and evaluated using some algorithms someone used in competition. Also referring to their another work next Monday: [26]

2.2.4 [27] RankME: reliable human ratings for natural language generation

- Problem of human rating for NLG: consistency, distinct criteria, relative assessment, etc.
- Solution: rank-based Magnitude Estimation (RankME), with relative ranking on continuous scale.
- How to assess the rating? Intra-class correlation coefficient (ICC)

2.2.5 [25] Using aspect extraction approaches to generate review summaries and user profiles (Airbnb)

- Task: Aspect extraction. Subtasks: (1) extract a representative sentence from a set of listing-specific reviews for a number of pre-defined aspects (e.g: cleanliness, location). (2) The suitability of aspect embeddings to represet guest profiles.
- Comparison between KMeans and ABAE (Attention-based aspect extraction. He et al., 2017), both of which are much better than LDA in these aspect extraction tasks.

2.3 Machine Learning 1

Empire A

2.3.1 [29] Zero-shot sequence labeling: transferring knowledge from sentences to tokens

- Task: give each token in a sentence a label (of what?), without telling the model how to predict
- Previous work to visualize LSTMs using e.g., attention weights, usually work on only a few data samples, and qualitatively.
- Method: First train a word LSTM (with attentions) on the classification task (e.g. the uncertainty prediction task), and that attentions show which tokens are the most important. This is the golden annotation y. Also the supervised learning "upper bound" baseline.
- Where does the zero-shot learning come from? Given this trained network, perform a "back-prop from pseudo-label" operation, assuming the pseudo-label is 0. Calculate the gradients at the words. For those whose labels are already 0, the gradients shall be small. Those words labeled as 1 should have large gradients. In this paper, this threshold is set to 1.5 deviation.

Takeaway The evaluation should be those of the supervised classifier's accuracies – zero-shot learning can *not* give this kind of per word accuracies. But visualizing the magnitude of gradients is a good idea to visualize LSTMs.

2.4 Machine Learning 2

Empire A

2.4.1 [5] Deep dirichlet multinomial regression

- Topic models. e.g. supervised topic models. What if the etadata are high-dimensional, structured, or may not directly relevant to modeling topics.
- Backbone: LDA. Change to DMR: sample from document-specific priors.
- From DMR to deep DMR
- Trained with Gibbs sampling.

2.4.2 [32] Training structured prediction energy networks with indirect supervision

- Structured prediction
- Parameterize energy function over y as a DNN -> can find the min of E using gradient descent.
- Supervised learning: Structured SVM (Belanger and McCallum, 2016)
- Indirect supervision
- Rank-based training

2.4.3 [15] Anchored correlation explanation: topic modeling with minimal Domain Knowledge

- How to do topic modeling with thousands of information bottleneck?
- LDA is a a generative topic model. Goods and bads of generative modelings.
- Topic model that learns topics through information-theoretic criteria.
- CorEx (total Correlation Explanation): a topic is a binary latent factor. Goal: find factors that make words conditionally independent.

$$\min_Y TC(W_1..W_n|Y) = \min_Y D_{KL}\left(p(w_1,..w_n|y)||\Pi_i p(w_y|y)\right)$$

  $TC(W|Y) = 0$ iff the topics "explain" all the dependence (total correlation). Here comes the Correlation Explanation name.

- Then rewrite the objective as mutual information

$$\min_Y \Sigma_i I(Y_j : W) - \Sigma_i \alpha_{i,j} I(W_i; Y_j)$$

  where $\alpha_{i,j} = \frac{I(W_i:W_j|Y_{k\neq j})}{I(W_i:Y_j)}$, which is the unique information in Y_j about W_i

  Then transform a combinatorial to a continuous optimization

- Extensions: (1) hierarchical CorEx; (2) semi-supervised learning.
- Anchored CorEx objective is exactly maximizing the information bottleneck.

2.4.4 [42] Aspect-augmented adversarial networks for domain adaptation

- Problem: Transfer learning, but both source and target classifiers operate over the same domain.
- Method: Learn a document-level representation that is hard to tell the domain, but easy to tell the class label.
- The encoder contains: A CNN per sentence; Improve adversarial training by reconstruction.
- Apply relevance score using a small set of keyword rules.

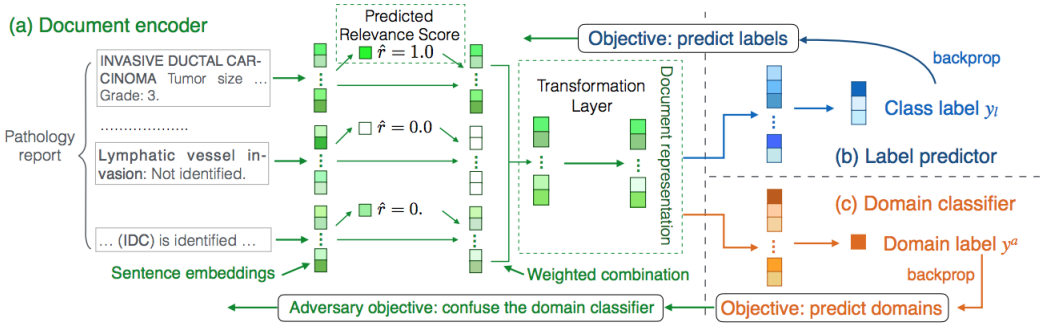


Figure 2: Aspect-augmented adversarial network for transfer learning. The model is composed of (a) an aspect-driven document encoder, (b) a label predictor and (c) a domain classifier.

2.5 SRW highlights

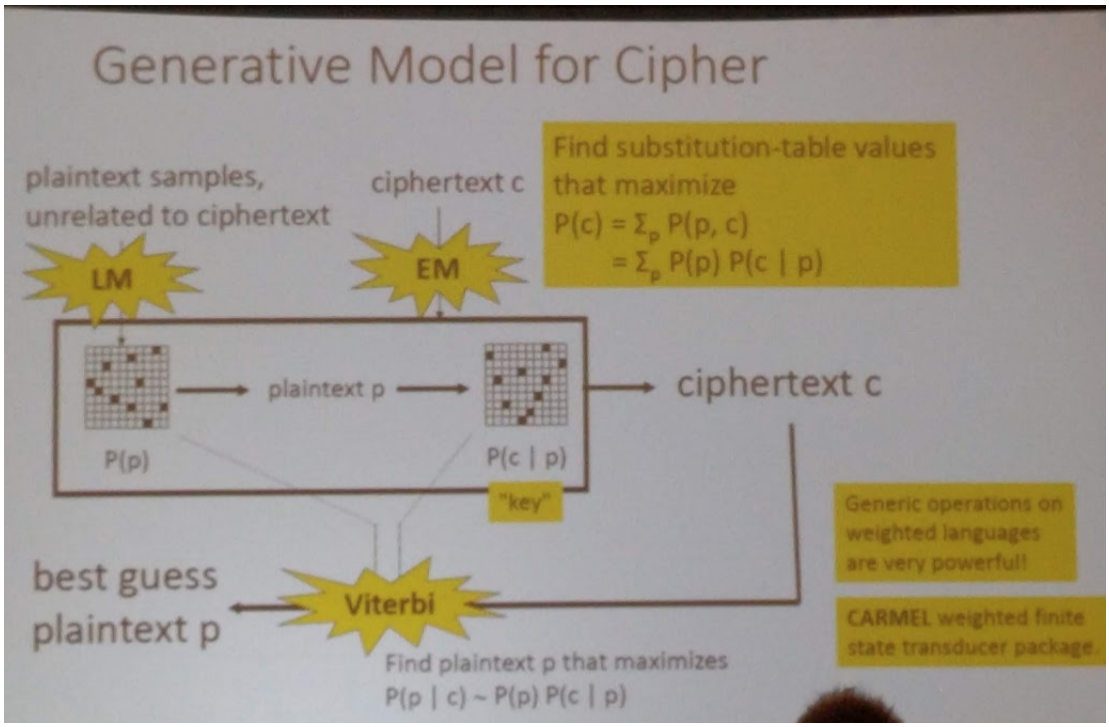
Empire C

2.5.1 [12] Igbo diactritic restoration using embedding models

- Igbo language: more spoken than written, and low-resource for NLP. (mostly south-eastern Nigeria)
- Problem: diacritic ambiguity (same wordkey, but different meanings)
- Embedding projection: align English embedding to Igbo language, using an alignment dictionary.
- Diacritic retoration proces: during evaluating candidate instances, choose the one with the maximum (cosine?) similarity in the embedded vector.

2.5.2 [1] Towards generating personalized hospitalization summaries

- Problem: summarization
- Method: FIrst build concept graph via UMLs, extract physician / nursing concepts to include. Then simplify. Then Arrange event ordering.



2.5.3 [37] Alignment, acceptance, and rejection of group identities in online political discourse

- Trump and Clinton supporters tend to use and align pronouns differently.
- Their rhetorical words are not substantially different.

3 Sunday 0603

3.1 Morning Keynote: The moment when the future fell asleep

Professor Kevin Knight

Decipher Deciphering of some ancient languages.
Decipherment is the original NLP problem.

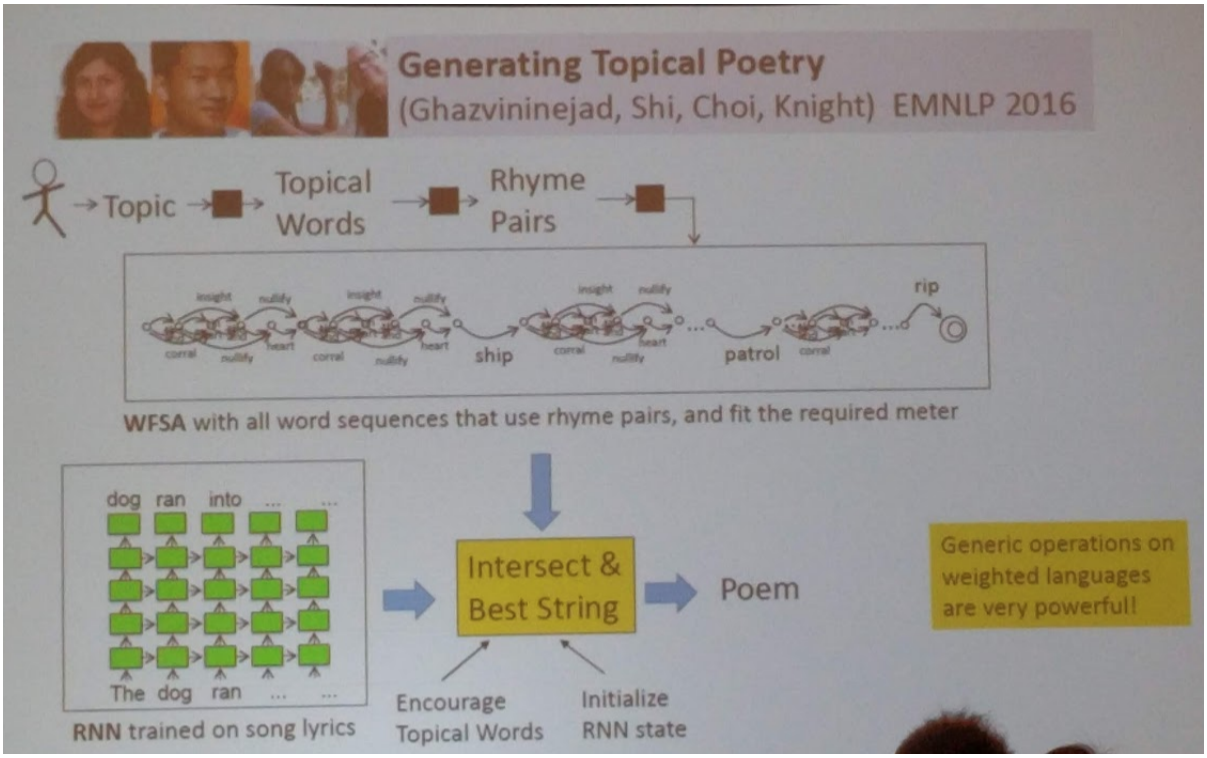
Generative model for cipher

Recent works

- Pixel image -> OCR + decipher in the system -> plain text. No supervision segment and clustering. After that, apply noisy-channel methods with plaintext language model.
- Zodiac cihpers: Z340, Z32, Z13
- Improve on the plaintext language models might lead to lower decipher error.

Poetry generation

- Can generate poets e.g.: Ghazycininejad, Shi, Choi, Knight EMNLP 2016
- Hafez: an interactive poetry generation system (Ghazvininejad et al. ACL demo 2017)



RNNs for storytelling

- How to memorize a random 60-bit string?
- RNNs as weighted language recognizers [8]
- Does string-based NMT learn source syntax? (EMNLP 2016)
- Why neural translations are the right length? (EMNLP 2016).
- Towards controllable story generation? (Peng et al., NAACL 2018 storytelling workshop)
- Paper anstract writing through abstract mechanism (Want et al, ACL 2018)
- Neural poetry translation [17]

Why do automatic outputs look so different (to us) than what was trained on?

Conclusions Still a long way to go. NLP for entertainment, commerce.

3.2 Morning Posters

3.2.1 FEVER: a large scale dataset for fact extraction and verification [38]

- A dataset containing 185,000 facts. Each wiki passage (?) has (humanly labeled) several claims (true or false).

3.2.2 Efficient sequence learning with group recurrent networks [16]

- Sequence learning task.
- Divide the hidden layer into two parts, so that computation is more efficient.
- Mix two parts of the hidden representation by the rearrange() function, so that the inter-group correlation can be considered more.

3.2.3 Embedding syntax and semantics of prepositions via tensor decomposition [18]

- Train a word embedding considering the tensor decomposition and preposition embeddings.
- The loss function contains ALS model term and the bias scalar parameters.

3.2.4 Semi-supervised event extraction with paraphrase clusters [13]

- First cluster the articles. Then identify "easy" events (after running a pre-trained supervised system on all sentences). Select most likely triggers for "hard" mentions.

3.3 Machine Learning 3

Empire A

3.3.1 Deep generative model for joint alignment and word representation [31]

- Optimize the variational lower bound of the marginal likelihood of a sentence pair: $P_\theta(x_i^m, y_j^n|m, n)$ where x_i^m and y_j^n are word observations in languages 1 and 2 respectively.

3.3.2 Evaluating the stability of embedding-based word similarities

- Cosine similarities are not satable. They have biases w.r.t corpus from which the word2vec are trained.
- Question: what do embeddings represent? They measure the properties of a curated corpus, not the word themselves.
- Two views of embeddings: downstream-centered or corpus-centered. (This work focus on the corpus-centered view)
- The embeddings are calculated using LSA (latent semantic analysis + tf-idf using `sklearn`), SGNS (skip-gram with negative sampling), GloVe, and PPMI (positive point-wise mutual information). Each methods contain fixed, shuffled, and bootstrap settings.
- Measure the cosine similarity bound of 20 query words. Those with lower variances is more stable.

Takeaway Should study from the methodology for designing experiments.

3.3.3 Learning word embeddings for low-resource language PU learning [21]

- Problem: large datasets are required to train datasets. Might be hard to find this large of corpus for low-resource languages. (this project focus on those with low-resource, but not those very-low-resource languages)
- Problem: Sparsity of the co-occurrence matrix (>99% of them are zero). Can be true zeros or missing entries (can co-occur but just has not in the given corpus)
- Motivation: word2vec use negative sampling, which only subsamples for some of the not mentioned.
- Propose a PU-learning framework for training word embedding. The learning alg deals wiht all negative pairs.
- Three components in the framework: (1) Pre-processing (building the o-occurrence matrix -> scale counts by PPMI metric w.r.t [Levy '15])

  (2) PU-learning for matrix actorization: $A \approx W^T H$, where we try to optimize (using coordinate descent) (see paper for the equations)

  (3) Post-proessing. Average w_i^T and h_i to get the resulting word vector.

3.4 Afternoon keynote: building innovative startups, products, and services – personal insights

Daniel Marcu (Amazon)

- Hard to annotate all domains – they are just too many of them. Very important therefore to enable domain adaptation.
- Commercial requirements usually forces us to be short-sighted.
- The most important lesson: We owe success to those people we worked with.
- Example of exploration in NMT structures.
- Important to know that the world is not just the pinnacle we focus on (e.g., in PhD). The world is the whole circle (big picture) – much more than what you have been focusing on pushing.
- Q: expectation between tech people and marketing people. People might like to go hype; also scientists should not make overly promising claims.
- Some tasks are hard to evaluate. These will be what a lot of future projects work on. e.g: quality of Alexa communication.

3.5 Afternoon Posters

3.5.1 Diverse few-shot text classification with multiple metrics [41]

- Task: Few-shot learning in diverse tasks.
- Propose an adaptive metric learning approach that automatically determines the best weighted combination from a set of metrics obtained from meta-training tasks for a (new) few-shot task.
- Matrix-completion based task clustering

3.5.2 Cross-lingual learning-to-rank with shared representations [33]

- Task: For each query-document pair, learn a mapping. More specifically, a query CNN and a document CNN compress the query and document into a hidden embedding.
- Several models to learn transferrable knowledge between languages. A basic "cosine model" minimizes the cosine similarity between the query and document embeddings. This does not work well on low-resource languages.
- A deep model adds a MLP to learn a similarity score given the embedding. How to make this model work on low-resource setting?
- Parameter-sharing is the improvement. Use the query CNN and the MLP trained on a high-resource language, and fine-tune using a low-resource language.

3.5.3 Are all languages equally hard to language-model? [10]

- Hypothesis: inflectional morphology makes a language hard to model. LM performance negatively correlated with morphological counting complexity.
- Correlation disappears when modeling lemmata instead of forms.
- Different languages contain varying bits per (English) character.
- The comparison methods are the takeaways at the end of the day.

3.6 Text Mining 1

3.6.1 Explainable prediction of medical codes from clinical text

- Task: the clinical coding problem
- Model: word embed -> CNN features -> attend (one for each label) -> classifier (Logistic Regression)

3.6.2 Event-time extraction with a decision tree of neural classifiers [30]

- Temporal links annotations
- Problem: sparse annotation of event times.
- TLINK annotation: dense annotation of event times. (ACL '16)
- Temporal anchoring of events given complete documents.
- Dataset: TimeBank-EventTime Corpus

3.7 Test of Time

Empire B

3.7.1 Remembrance of Aravind Joshi

- 1929 - 2017
- Centering: a framework for modeling the local coherence of discourse; the penn discourse treebank; tree adjoining grammars.

3.7.2 BLEU

- ACL 2002, Kishore et al.
- "Bilingual Evaluation Understudy": compare short runs of candidate text against reference translations.
- Not necessary to match pair-wise
- All words are equally important: all you need is a tokenizer. Set up a brevity penalty BP:

$$BP = \begin{cases} 1 \text{if c} > \text{r} \\ e^{1-r/c} \text{otherwise} \end{cases}$$

- Modified u-gram precision: average log with unit weights.
- BLUE = $BP \times exp(\sum_n w_n log p_n)$ where p_n is the n-gram precision, and positive weights w_n sum up to 1.
- Stunningly simple, surprisingly simple.

Retrospective

- Context: DARPA: slow and expensive for human evaluations; long pause in funding.
- Hard to sell in 2002: "quantity leads to quality". "Don't attempt to divine human judgment for every sentence. Rather, average out individual sentence judgment errors over a corpus."
- Polarization: people have polarized reviews towards BLEU.
- It is an understudy – never meant to replace human judgments.
- Q: criticism that BLEU penalizes stylicism? A: wow the metrics actually understand language.

3.7.3 Structured Perceptron

- EMNLP 2002, "Discriminative Training Methods for HMM", Michael Collins
- Background: structured prediction problems
- Dominant approach in the 1990s to do structured prediction as density estimation: model $p(x, y)$ or $p(y|x)$. e/g: Log-linear history-based models.

- Method: POS Tagging with HMMs
- Generalization bounds. (written in a PAC-style).
- Conclude: thoughts about learning and search.

  Approach 1: global training. e.g: CRF log-loss, max-margin

  Approach 2: Local training, etc., seq2seq
- Some hypotheses in 2002: (1) search is necessary for some structured prediction problems. (2) If you believe search is beneficial, then local normalization does not work. (3) Generalization bounds will be important in a scientific understanding of why / how machines learn. (see Dziugaite and Roy on PAC-Bayes for NN)

  Interesting to think about them.

3.7.4 A sentiment Odyssey

- EMNLP 2002, Thumbs up? Sentiment classification using ML techniques by Bo Pang et al.
- Background: people start to use internet increasingly prevalently. Then sentiment analysis is increasingly important. Also the dataset sizes were pretty small.
- Released the imdb movie review dataset.
- Inspirational slide? Do something interesting but that might not be deemed interesting as considered by other people. Since most people here have done PhD they should know how to get out of it.

4 Monday 0604

4.1 Keynote: Google assistant or my assistant?

Background Early conversation systems occurred in 1990s. Nowadays, however, the conversation systems stiil need a long way to go, because they should be multi-domained, and scalable (a lot of data-hungry methods are hard to scale).

4.1.1 Task-oriented dialogue as a collaborative game

A seeker and a provider agents. Human are usually the seeker, but machine could be seeker as well. Seeker has a goal but no access to solution. The provider has solutions but does not know the goal.

Provider language understanding, state tracking, {dialogue manager, response generator} that queries the backend (action provider / knowledge bases).

- Trained by RL aiming to optimize longer term dialogue reward (Levin et al, IEEE TASLP 2000, Singh et al AAAI 2000)
- Hierarchical RL for multi-domain dialogues (Peng et al, EMNLP 2017)
- Reward estimation (Su et al, ACL 2016)

Seeker language understanding, dialogue state tracking, {dialogue manager, response generator} that interacts with the scenario.

- usually termed as "user simulators".
- e.g: seq-to-seq models;
- probabilistic, agenda based [35]

Crawl: machines talking to machines [36] tries to produce dialogues that make sense to humans.

4.1.2 Dialogue system components attempts to solve challenges

Conversational Language Understanding LSTM + attention?

How to Integrate conversation context? Context vector, attention over history, time decay attention (Su et al, 2018 NAACL)

Challenge: scaling CLU to new verticals.

- Zero-shot learning using slot tag and description embeddings as additional input during parsing (Bapna et al, Interspeech 2017)
- Train on target (LeFevre et al, Interspeech 2010)
- Test on source (Jabaian et al, ICASSP 2011)
- Joint learning on source and target (He et al, ICASSP 2013)
- Cross-lingual embeddings (Upadhyay et al, ICASSP 2018)

Referring expression resolution for situated dialogues

Dialogue state tracking Agent's estimate of the user's goal(s) based on the dialogue history. Research on DST has been fostered by the dialogue state tracking challenges:

- Delexicalised RNNs
- Neural belief tracker
- End-to-end memory networks for DST
- Recent pointer networks (Xu and Hu, arXiv 2018)
- Simulated datasets focus on entities that were not previously observed
- (Rastogi et al, IEEE ASRU 2017) Focus only on the relevant set of slot values

Accuracy and scalability are important, but efficiency is important too! Hierarchical recurrent neural network (Gupta et al, INterspeech 2018) (Rastogi et al, SigDial 2018) that halved the parameter.

4.1.3 End-to-end learning

e2e dialogue with deep RL

- Component-wise training benefits from additional data for each component
- Supervised learning (Li et al ICASSP 2017, Bordes et al ICLR 2017)
- RL (Williams et al ACL 2017, Dhingra ACL 2017)
- Eng-to-end dialogue models with human teaching (poster at NAACL 2018)

4.1.4 Other interesting topics

- Diverse in-domain dialogue data
- scaling to new domains / languages
- integrating context
- scalable multi-domain state tracking
- learning about users
- Understanding and tracking with complex / compositional representations
- Generating multi-modal content
- Situated, multi-modal dialogues
- Latent understanding

4.2 Morning posters

4.2.1 Deconfounded Lexicon induction for interpretable social science [28]

- Need to mae social models transparent and interpretable. Formalize as" induce a lexicon that is predictive of a set of target variables yet uncorrelated to a set of confounding variables.

  Formally, look for L that maximizes informativeness coefficient:

  $$\mathcal{I}(L) = \mathbb{E}[Var[\mathbb{E}[]Y|L(T), C|C]]$$

- So they are monitoring the coefficient, but not optimizing it. The optimization is through an adversarial selector.
- Two deep learning algorithms for this task.

Takeaway The induced lexicons reminds me of anchor variables (Halpern et al)

4.2.2 Learning to rank Q-A pairs using hierarchical RNN with latent domain clustering

- Train hierarchical RNN, then cluster the latent dimensions
- Then incorporate each latent vector by the *similarity* of it and all cluster topic vectors.
- Interestingly, the latent topic cluster does not agree with the human clustering. I think this is reasonable – you can't ensure that neural networks learn things exactly in the euclidean space.

4.2.3 Supervised and Unsupervised transfer learning for question answering [9]

- Model: memory networks. Use CNN to process the embeddings.
- For the supervised part: fine tune the upper network layers, or the whole networks, even including the embedding networks.
- For the unsupervised part: self-train bootstrapping.
- Dataset: Multiple choice QA.

Takeaway This is an example where direct fine-tune transfer learning works. It is understandable that multiple choice QA knowledge are easy to transfer than other tasks like translation or classification.

4.2.4 Deep communicating agents for abtractive summarization[7]

- Agent model: pointer network (seq2seq) for abstractive summarization.
- One paragraph per agent. Break down the difficult long text reading problem down to multiple agents.
- MLE loss: $L_{MLE} = -\sum_t logp(y_t|t_{1..t-1}, d)$. Try to make the predicted next word as close to the next ground truth word as possible.
- Semantic cohesion loss: $L_{sem} = \sum_q cos(s_q, s_{q-1})$. This encourages the generated sentences to be coherent with the previous sentence.
- Also contains an RL loss: $L_{RL} = (r(\tilde{y}) - r(\hat{y}))\sum_t logp(\hat{y}_t|\hat{y}_{1..(t-1),d})$. How are the reward decided? Self-critical approach. It is calculated by comparing the similarity between generated sentence and the ground truth.
- Putting the loss together: $L_{mixed} = \gamma L_{RL} + (1-\gamma)L_{MLE}$
- Evaluate using incremental ROUGE scores, considering the intermediate rewards.

Takeaway So in many NLP works the reinforcement learning rewards refer to the one calculated upon reaching the end of a passage. In robotics setting RL loss corresponds to the end of an epoch. Note this difference.

4.2.5 Key2Vec [24]

- Procedure: Fasttext-skipgram

4.2.6 Unsupervised keyphrase extraction with multipartite graphs [6]

- A multipartite graph: nodes from each group does not connect to each other.
- Adjust the weights of graph edges by some metrics w.r.t the entries.

4.2.7 Estimating summary quality with pairwise preferences [43]

- Set up generated preferences as games: the better sentences are more likely to win the game.
- Compute Bradley-Terry scores: $p(s_x > s_y) = \frac{u(s_x)}{u(s_x)+u(s_y)}$. Now you have a list of preferences.
- Value of a summary (collection of sentences): $v(S) = \sum_i^{|S|} w_i u(s_i)$
- Automatically generating preferences (refer to paper for details)
- Using additional automatically generated labels can further improve accuracy.

4.2.8 Which scores to predict in sentence regression for text summarization?[44]

- Recall is biased towards long sentences.
- Ordering according to prevision leads to better summaries.
- Takeaway: better select sentences according to ROGUE precision in summarization tasks.

4.3 Generation 3

Empire C

4.3.1 Interpretable charge predictions for criminal cases

- Learning to generate court views from fact descriptions [40]
- Overview: charge prediction. Input: fact description in a criminal case. Output: charge label (e.g. drunk driving, intentional injury, etc.)
- Previous work lack interpretability
- What is court view? A written explanation from judges to interprete the charge, including rationale and charge label (only rationale generation this paper).
- High quality rationale? (1) Should contain relevant details; (2) should be charge-discriminative.
- Model: label-conditional seq2seq model. Bi-LSTM with attention.
- How is the label-conditional work? The label is predicted using encoder attentions. Then the predicted label is passed as an additional input into every step of the decoder.

4.3.2 Delete, Retrieve, Generate: a simple approach to sentiment and style transfer [23]

- Text attribute transfer
- Previous work example: adversarial content saparation: (Shen et al 2017, Fu et al 2018)
- Proposed method: (based on seq2seq)

  (1) Delete the words most indicative to the sentiment (see paper for details)

  (2) Retrieve: decide what words to insert into context.

  (3) Generate:
- Comparison models: TemplateBased, DeleteOnly, DeleteAndRetrieve

4.3.3 Adversarial example generation with syntactically controlled paraphrase networks [20]

- Adversarial examples generation for images (Goodfellow 2015 ICLR)
- Lexical adversaries (ACL 2018) or syntactic adversary
- Introduce ML approach for syntactic adversary generation
- SCPN: (1) acquire sentential paraphrase pairs through neural backtranslation (using ParaNMT corpus, ACL 2018); The sentences translated back have some syntactic differences. These are uncontrolled paraphrases.

  Syntactic parse both sentences.

  (2) automatically label with Stanford parser.

  (3) Copy mechanism on encoder (of the LSTM seq2seq model)
- At test time, use only the top two levels of the parsers.
- Evaluation: (1) intrinsic evaluations; (2) adversarial evaluation (sentiment analysis with Stanford Sentiment Treebank).
- Takeaways: (1) SCPN paraphrases does not lose paraphrase quality in comparison to NMT-BT baseline. (2) Adversarial evaluation: 30% - 40% broken (about twice as many as NMT-BT)
- How to make the models more robust w.r.t this kind of adversarial attack? Include paraphrased sentences into the training sets of classifier. This can be helpful against adv attack

4.4 Sentiment Anaysis 2

2pm Empire A

4.4.1 Sentiment analysis: it's hard[22]

- Dataset: MTSA: McGill Twitter Sentiment Analysis (7026 tweets). People disagree with some tweets.
- Don't purge the data (bring in noise), or purge (lose some data)
- Bring in a "complicated" label.
- Still 7.9% of tweets do not have agreed labels.
- Is that noisy annotators or data are qualitatively distinct? -> data that are hard to classifiers are also hard for annotators to label.
- Can we detect "complicated" data? (Not yet)
- Perspective: raw annotations may offer more informative signal for classifiers.

Takeaway Before questioning on the classifiers, we should also look at how the data are labeled. Also this reminds me of Hinton's dark knowledge.

4.4.2 Multitask learning of pairwise sequence classification tasks over disparate label spaces[2]

- Task: MTL
- Multi-task learning network: shared hidden layers + one output layer per task
- Model 1: label embedding layer. Labels for all tasks are embedded in a separate space. This can enable us to learn the relationships between the labels in the joint embedding space.
- Model 2: label transferring network. Learn to produce pseudo labels for target task.
- Model 3: semi-supervised MTL with LTN.

4.4.3 Human needs categorization of affective events using labeled and unlabeled data[11]

- Affective events: why are they positive / negative? Try to explain them with human needs.
- Human needs: physiological, health & safety, leisure & aesthetic, finance, social, cognition & education, freedom & accessibility, { emotions / sentiments, optinions (misc. categories)}, and other.
- Hypotheses: (1) the polarities of affective events often stem from whether experiencers' human needs are satisfied or violated. (2) most affective events can be explained by a small number of human needs.
- Dataset: annotators had pretty good agreements kappa (pairwise agreeent scores)
- Models: event expression classifier and event context classification.
- Semi-supervised algorithms:

  (1) self-training: event expression classifier; iteratively retrain.

  (2) a expression classifier and context classifier can be set up for a co-training framework.

4.4.4 Multimodal emoji prediction [4]

- Emojis are powerful multimodal communication forms.
- Task: given the text and image, predict the emoji included in the text.
- Relevant work: (EACL 2017 "Are emojis predictable?") ("FastText")
- Model: FastText on text, ResNet-101 on image, freeze the layers, train a Logistic Regressor on top of them. Simple model.
- Multimodal dataset: Instagram

4.5 Outstanding paper session

Empire

4.5.1 Deep contextualized word representations Best paper award

- Language understanding needs context.
- Propose ELMo: EMbeddings from language Models.
- What is ELMo?

  Compute contextual vector: $c_k = f(w_k|w_1...w_k..w_n)$

  $ELMo = \lambda_2 R_2 + \lambda_1 R_1 + \lambda_0 R_0$ where R_k is the k^{th} layer of the LSTM. Specifically k=0 corresponds to the input (word embedding) and k=1, 2, ... are the hidden layers output.
- Properties of ELMo representations: (1) unsupervised; (2) contextual; (3) deep; (4) character-based; (5) versatile.
- 4GPU, 2 weeks to train the model.

4.5.2 Neural text generation in stories using entity representations as context

- Can we use entity representations as a form of context to improve text generation for stories?
- Three evaluations: mention generation, (pairwise) sentence selection, human evaluation
- base model: seq2seq with attention. (but it usually does not mention entities in previous sentences)
- Want the mentions also be carried out in a natural way.
- Full model: Each step contains the content generated in current sentence and previous sentence, and the current state of the entities entioned in the docuent so far.
- Dataset: Toronto book corpus: adventure books (containing entity annotations)
- Future directions: deeper entity knowledge: social commensense, modeling inter-entity relationships. Structure in story generation. New domains: news articles, recipes. etc.

4.5.3 RNNs as weighted language recognizers [8]

- Strings and probabilities. RNNs are probabilitstic automatas.
- Formal properties of RNNs:

  For any turing Machine, can construct an RNN to simulate it. (Siegelmann and Sontag 1995)

  No thinking time between action and inputs.
- Questions about RNN's formal properties: consistency; best string; equivalence; minimization.
- Consistency: $\sum_s R(s) = 1$?

  Problem of inconsistent PCFG.

  However, PCFG trained from EM are consistent

  Consistent or inconsistent RNN. Empirically as SGD training proceeds, RNNs become more consistent
- Highest=weighted string?

  Highest-weighted string under lenght bound is NP-complete.
- Equivalence? No.
- Minimization? Undecidable.

Takeaway 1. Quite surprising to see such a formal (theoretical) analysis on RNN in NAACL (instead of COLT).
2. The first author did this during a summer internship. She entered university in the same year (2014) as me. What am I doing orz...

5 What do people work on?

NAACL 2018 is the first ML / NLP conference I go to (went to ICRA 2017 but that was robotics), and I am looking for directions for future thesis / projects / etc., so this is a natural question to ask. In general, their topics vary, but some common trends could be observed.

First trend is **new progresses on traditional tasks** (sequential prediction, summarization, comprehension, translation, question-answering, NER, etc.) Word embedding, for example, is another direction that is continuously improved upon. There are word embeddings that incorporate language models (the best paper this year), incorporate prepositions, etc.

Second trend is the emerging of **new tasks, inspired by real-world applications**. Data-efficient learning (few-shot learning, semi-supervised learning, etc.) algorithms have been applied, for example. Technologies relating to health (e.g., EHR records, CLPsych) and society (fairness, law, etc.) are also mentioned. In business, Alexa and Google Assistant require and inspire advances in conversational agents. Text mining from travel / hotel / product reviews inspire some very interesting works.

Third is **new tasks brought in by the advance of core machine learning**. CNNs (even Transformers) have started to take the place of RNN for sequential tagging tasks. Many domain adaptation papers begin to use shared-private layers and adversarial loss. Reinforcement Learning are becoming popular as well.

Fourth trend is the increased focus on **developing explainable models and evaluation metrics**. There was a talk on the stability of word embeddings, several posters addressing the evaluation metrics. To make the models more explainable, some tools in information theory are applied (e.g: information bottleneck for topic modeling in TACL).

References

[1] Sabita Acharya, Barbara Di Eugenio, Andrew Boyd, Richard Cameron, Karen Dunn Lopez, Pamela Martyn-Nemeth, Carolyn Dickens, and Amer Ardati. Towards generating personalized hospitalization summaries. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 74–82. Association for Computational Linguistics, 2018.

[2] Isabelle Augenstein, Sebastian Ruder, and Anders Søgaard. Multi-task learning of pairwise sequence classification tasks over disparate label spaces. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1896–1906. Association for Computational Linguistics, 2018.

[3] Ramy Baly, Mitra Mohtarami, James Glass, Lluís Màrquez, Alessandro Moschitti, and Preslav Nakov. Integrating stance detection and fact checking in a unified corpus. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 21–27. Association for Computational Linguistics, 2018.

[4] Francesco Barbieri, Miguel Ballesteros, Francesco Ronzano, and Horacio Saggion. Multimodal emoji prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 679–686. Association for Computational Linguistics, 2018.

[5] Adrian Benton and Mark Dredze. Deep dirichlet multinomial regression. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 365–374. Association for Computational Linguistics, 2018.

[6] Florian Boudin. Unsupervised keyphrase extraction with multipartite graphs. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 667–672. Association for Computational Linguistics, 2018.

[7] Asli Celikyilmaz, Antoine Bosselut, Xiaodong He, and Yejin Choi. Deep communicating agents for abstractive summarization. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1662–1675. Association for Computational Linguistics, 2018.

[8] Yining Chen, Sorcha Gilroy, Andreas Maletti, Jonathan May, and Kevin Knight. Recurrent neural networks as weighted language recognizers. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2261–2271. Association for Computational Linguistics, 2018.

[9] Yu-An Chung, Hung-yi Lee, and James Glass. Supervised and unsupervised transfer learning for question answering. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1585–1594. Association for Computational Linguistics, 2018.

[10] Ryan Cotterell, Sebastian J. Mielke, Jason Eisner, and Brian Roark. Are all languages equally hard to language-model? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 536–541. Association for Computational Linguistics, 2018.

[11] Haibo Ding and Ellen Riloff. Human needs categorization of affective events using labeled and unlabeled data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1919–1929. Association for Computational Linguistics, 2018.

[12] Ignatius Ezeani, Mark Hepple, Ikechukwu Onyenwe, and Enemouh Chioma. Igbo diacritic restoration using embedding models. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 54–60. Association for Computational Linguistics, 2018.

[13] James Ferguson, Colin Lockard, Daniel Weld, and Hannaneh Hajishirzi. Semi-supervised event extraction with paraphrase clusters. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 359–364. Association for Computational Linguistics, 2018.

[14] Yao Fu and Yansong Feng. Natural answer generation with heterogeneous memory. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 185–195. Association for Computational Linguistics, 2018.

[15] Ryan J Gallagher, Kyle Reing, David Kale, and Greg Ver Steeg. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *arXiv preprint arXiv:1611.10277*, 2016.

[16] Fei Gao, Lijun Wu, Li Zhao, Tao Qin, Xueqi Cheng, and Tie-Yan Liu. Efficient sequence learning with group recurrent networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 799–808. Association for Computational Linguistics, 2018.

[17] Marjan Ghazvininejad, Yejin Choi, and Kevin Knight. Neural poetry translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 67–71. Association for Computational Linguistics, 2018.

[18] Hongyu Gong, Suma Bhat, and Pramod Viswanath. Embedding syntax and semantics of prepositions via tensor decomposition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 896–906. Association for Computational Linguistics, 2018.

[19] Pankaj Gupta, Benjamin Roth, and Hinrich Schütze. Joint bootstrapping machines for high confidence relation extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 26–36. Association for Computational Linguistics, 2018.

[20] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885. Association for Computational Linguistics, 2018.

[21] Chao Jiang, Hsiang-Fu Yu, Cho-Jui Hsieh, and Kai-Wei Chang. Learning word embeddings for low-resource languages by pu learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1024–1034. Association for Computational Linguistics, 2018.

[22] Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895. Association for Computational Linguistics, 2018.

[23] Juncen Li, Robin Jia, He He, and Percy Liang. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874. Association for Computational Linguistics, 2018.

[24] Debanjan Mahata, John Kuriakose, Rajiv Ratn Shah, and Roger Zimmermann. Key2vec: Automatic ranked keyphrase extraction from scientific articles using phrase embeddings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 634–639. Association for Computational Linguistics, 2018.

[25] Christopher Mitcheltree, Veronica Wharton, and Avneesh Saluja. Using aspect extraction approaches to generate review summaries and user profiles. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 68–75. Association for Computational Linguistics, 2018.

[26] Mitra Mohtarami, Ramy Baly, James Glass, Preslav Nakov, Lluís Màrquez, and Alessandro Moschitti. Automatic stance detection using end-to-end memory networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 767–776. Association for Computational Linguistics, 2018.

[27] Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. Rankme: Reliable human ratings for natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 72–78. Association for Computational Linguistics, 2018.

[28] Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wagner. Deconfounded lexicon induction for interpretable social science. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1615–1625. Association for Computational Linguistics, 2018.

[29] Marek Rei and Anders Søgaard. Zero-shot sequence labeling: Transferring knowledge from sentences to tokens. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 293–302. Association for Computational Linguistics, 2018.

[30] Nils Reimers, Nazanin Dehghani, and Iryna Gurevych. Event time extraction with a decision tree of neural classifiers. *Transactions of the Association for Computational Linguistics*, 6:77–89, 2018.

[31] Miguel Rios, Wilker Aziz, and Khalil Simaan. Deep generative model for joint alignment and word representation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1011–1023. Association for Computational Linguistics, 2018.

[32] Amirmohammad Rooshenas, Aishwarya Kamath, and Andrew McCallum. Training structured prediction energy networks with indirect supervision. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 130–135. Association for Computational Linguistics, 2018.

[33] Shota Sasaki, Shuo Sun, Shigehiko Schamoni, Kevin Duh, and Kentaro Inui. Cross-lingual learning-to-rank with shared representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 458–463. Association for Computational Linguistics, 2018.

[34] Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 35–41. Association for Computational Linguistics, 2018.

[35] Kashif Shah, Selcuk Kopru, and Jean David Ruvini. Neural network based extreme classification and similarity models for product matching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 8–15. Association for Computational Linguistics, 2018.

[36] Pararth Shah, Dilek Hakkani-Tur, Bing Liu, and Gokhan Tur. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 41–51. Association for Computational Linguistics, 2018.

[37] Hagyeong Shin and Gabriel Doyle. Alignment, acceptance, and rejection of group identities in online political discourse. In *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics: Student Research Workshop, pages 1–8. Association for Computational Linguistics, 2018.

[38] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819. Association for Computational Linguistics, 2018.

[39] Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. Label-aware double transfer learning for cross-specialty medical named entity recognition. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1–15. Association for Computational Linguistics, 2018.

[40] Hai Ye, Xin Jiang, Zhunchen Luo, and Wenhan Chao. Interpretable charge predictions for criminal cases: Learning to generate court views from fact descriptions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1854–1864. Association for Computational Linguistics, 2018.

[41] Mo Yu, Xiaoxiao Guo, Jinfeng Yi, Shiyu Chang, Saloni Potdar, Yu Cheng, Gerald Tesauro, Haoyu Wang, and Bowen Zhou. Diverse few-shot text classification with multiple metrics. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1206–1215. Association for Computational Linguistics, 2018.

[42] Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. Aspect-augmented adversarial networks for domain adaptation. *arXiv preprint arXiv:1701.00188*, 2017.

[43] Markus Zopf. Estimating summary quality with pairwise preferences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1687–1696. Association for Computational Linguistics, 2018.

[44] Markus Zopf, Eneldo Loza Mencía, and Johannes Fürnkranz. Which scores to predict in sentence regression for text summarization? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1782–1791. Association for Computational Linguistics, 2018.