Deep Learning for Social Media Text Mining (and beyond)

Ismini Lourentzou & ChengXiang Zhai {lourent2, czhai}@illinois.edu

University of Illinois at Urbana - Champaign, Computer Science Dept.

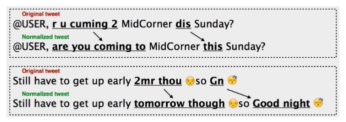
Co-authors in geolocation: Alex Morales, UIUC

Co-authors in relation extraction: Anna Lisa Gentile, Daniel Grul, Alfredo Alba, Steve Welch *(@IBM Almaden Research Center)*

Text Normalization

Text in twitter messages and other social media platforms often contains spelling errors, non-standard words, and acronyms.

- bridge communication issues and confusion across multiple groups
  - abbreviations and slang used by young people vs. older audience
  - different group dialects (e.g. African American vernacular)
- helpful pre-processing step for user-generated text
  - higher out-of-vocabulary (OOV) rates due to non-standard words
  - lower accuracy in NLP methods applied in social media (i.e. sentiment analysis, spam filtering, etc.)



correct spellings *rite → right*
expand abbreviations *tmrw → tomorrow*
phonetic substitutions *4eva → forever*

Word-level substitutions

- Create candidate replacements for each word ("generators")
  - word-level operations: capitalize, lowercase, smallest edit distance, google autocorrect, contractions (i.e. I'm → I am), data dictionary
- Learn the best substitution "generator"
  - pairs of feature vectors and corresponding best generator
  - minimum edit distance as metric for ranking generators

Sequence to Edits LSTM

- Create a dictionary mapping every word to a list of normalized forms
- Words with unique mapping are replaced *rite → right*
- Words with multiple mappings passed to LSTM *ur → {your, you are}*
- For every word with multiple mappings, calculate minimum-cost edit operations that covert an unnormalized word to its normalized version
  - character-level operations:
    delete, replace, input a character before the current index, none
- LSTM model trained on edit operations

ur → you are : $insert\_y\ insert\_o,\ insert\_\_\ insert\_a,\ insert\_e$

Category	1:1	1:N	N:1	Overall
Training	2,875	1,043	10	3,928
Test	2,024	704	10	2,738

ACL'15 WNUT Dataset[1]

Models	Precision	Recall	F-1
Word-Generator	0.7221	0.5897	0.6492
LSTM	0.9014	0.6829	0.7771

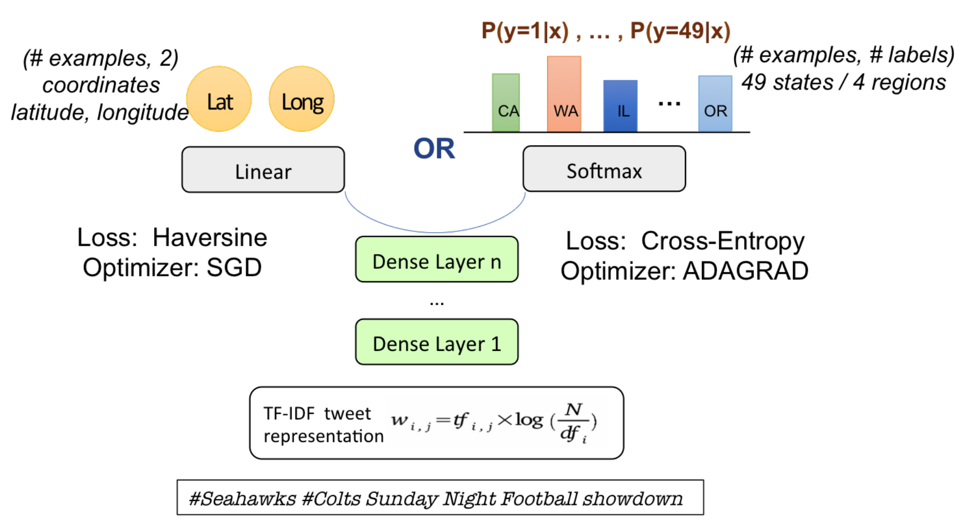
Results

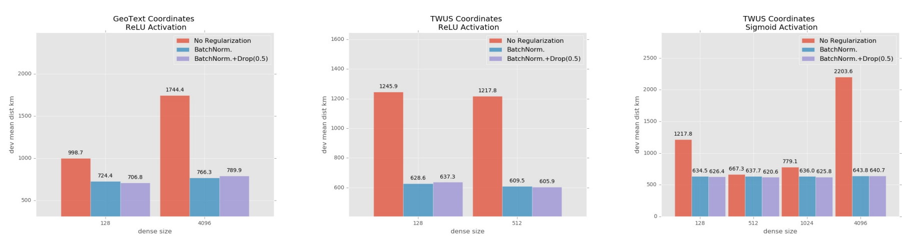
Text-based Geolocation Prediction

In this work, we study how to apply deep learning more effectively to solve the problem of text-based geotagging by systematically varying all the major decisions including the activation functions, layer and regularization choices with two different prediction task formulations

Dataset Name	Users	Sample Size	Region
GeoText	9.5K	380K tweets	Contiguous US
TwUS	450K	38M tweets	North America
TwWORLD	1.4M	12M tweets	English World Wide

Twitter Geolocation Datasets [2, 4, 3]





GeoText	States	Regions
Proposed method	44.3	67.3
Liu and Inkpen, 2015 (SDA)	34.8	61.1
Eisenstein et al., 2010 (Geo topic model)	24	58
Cha et al., 2015 (SC+all,word sequences)	41	67

Results on GeoText - classification (Accuracy)

GeoText	Mean	Median	Acc@161
Proposed method	747	448	29
Rahimi et al.,2017 (MDN-SHARED)	865	412	39
Liu and Inkpen, 2015 (SDA)	856	-	-
TWUS	**Mean**	**Median**	**Acc@161**
Proposed method	570	223	43
Rahimi et al.,2017 (MDN-SHARED)	655	216	42
Liu and Inkpen, 2015 (SDA)	733	377	24
TWWORLD	**Mean**	**Median**	**Acc@161**
Proposed method	1338	495	21
Wing and Baldridge (2014) & HierLR Unif	1715	490	33
Wing and Baldridge (2014) & HierLR k-d	1670	509	31

Results on regression (Error in km)

On-demand Relation Extraction

Extract relations of interest from free text.
Most NLP applications require domain-specific knowledge

- Which vitamins inhibit the absorption of other vitamins?
- Who is the biggest competitor of Apple?

Recent state of the art has been focusing on incorporating linguistic knowledge in (neural) architectures and maximizing performance by means of feature engineering. **Requisite: availability of large datasets**

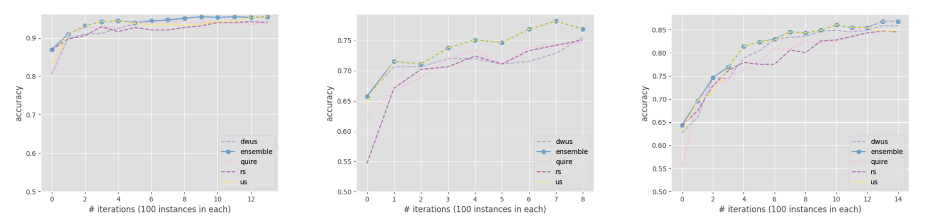
Unfeasible! The definition of a relation is highly dependent on the task at hand and on the view of the user

Ideally, we aim to achieve:

- fast training on any relation
- according to user-defined requirements
- under limited annotated data
- not relying on additional linguistic knowledge resources

Dataset	#examples	Relations
Semeval10 Task 8	10,717	9 types: Entity-Origin, Message-Topic, etc.
CausalADEs	1,420	causal drug-ADE relations from medical forum posts

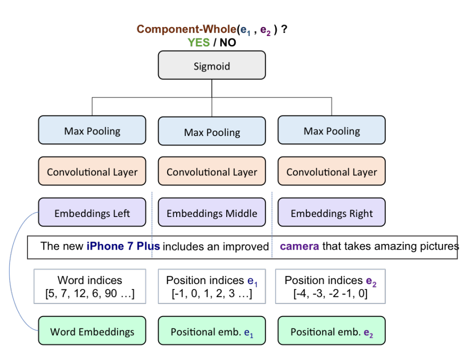
Neural models for on-demand relation extraction method with *human-in-the-loop*, starting from a few user-provided examples. Batch selection by identifying the best active learning strategy.



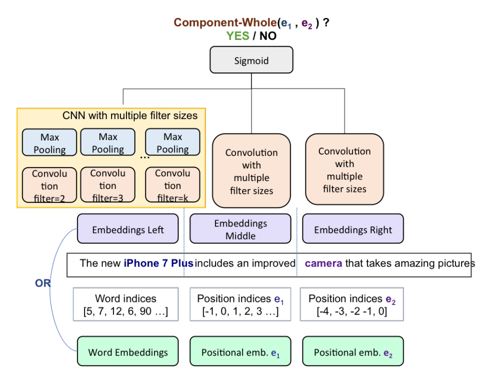
Member-Collection CNN context-wise split input | CausalADEs CNN context-wise split input | Entity-Origin CNN mult. filters - positional features

Sentence	y	$\hat{y}$	$P(\hat{y} = y \mid x)$
I was on **Crestor** for only two months when my knee just flared up in pain followed by **muscle pain**.	1	1	0.99
However, I am afraid to discontinue the **Paxil** due to fear of withdrawal symptoms and/or return of **panic attacks**	0	0	0.99
I felt like **Zoloft** turned me into a little bit of a **zombie**	1	0	0.722
I was **crying** at the drop of a hat until I started taking the **Celexa**, so has been a life saver in my opinion	0	1	0.497
put me on **prozac** and it made me more **jittery**	1	0	0.803

Examples of correct and incorrect predictions on CausalADEs



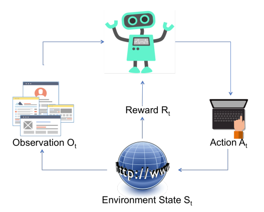
a) convolutional operations followed by maxpooling



c) convolutions with multiple filter sizes

Forthcoming Research

Reward Augmentation in Text-based Deep Reinforcement Learning



Emerging field of new applications of DRL, where the goal is NLU: recommend systems, query reformulation, web navigation, sentence simplification, web navigation, dialog systems etc. Many challenges studied from a linguistic perspective:

- **Sparse Rewards** guiding exploration, providing small and diverse "hints" that lead to higher rewards is crucial
- **Reward misspecification** in scenarios where human feedback is involved, we have to deal with inconsistencies and human errors, which can lead to a noisy reward function
- **External knowledge** some tasks are intuitive to humans, as they rely on knowledge of concepts and common sense, e.g. reasoning about entities and relations, a DRL agent shows poor convergence properties when directly trained by trial and error [5]. Supervised learning is leveraged to tackle this problem, but this solution comes at an additional computational cost and might not always be available.

References

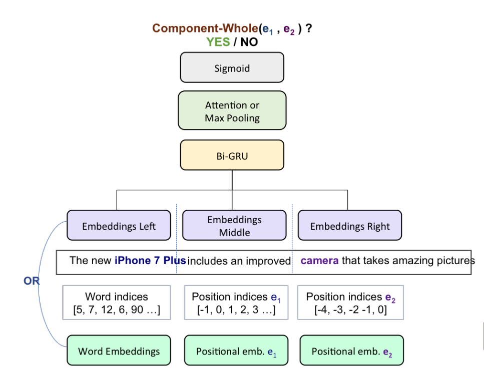
[1] T. Baldwin, M.-C. de Marneffe, B. Han, Y.-B. Kim, A. Ritter, and W. Xu. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In *Proceedings of the Workshop on Noisy User-generated Text*, 2015.

[2] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, 2010.

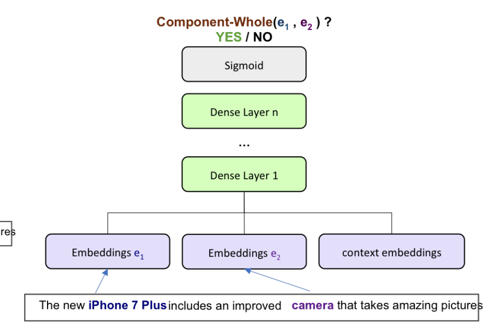
[3] B. Han, P. Cook, and T. Baldwin. Text-based twitter user geolocation prediction. *Journal of Artificial Intelligence Research*, 2014.

[4] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge. Supervised text-based geolocation using language models on an adaptive grid. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012.

[5] W. Xiong, T. Hoang, and W. Y. Wang. Deeppath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017.



b) bi-directional GRUs, followed by attention or temporal maxpooling



d) fully-connected multi-layer perceptron (MLP)