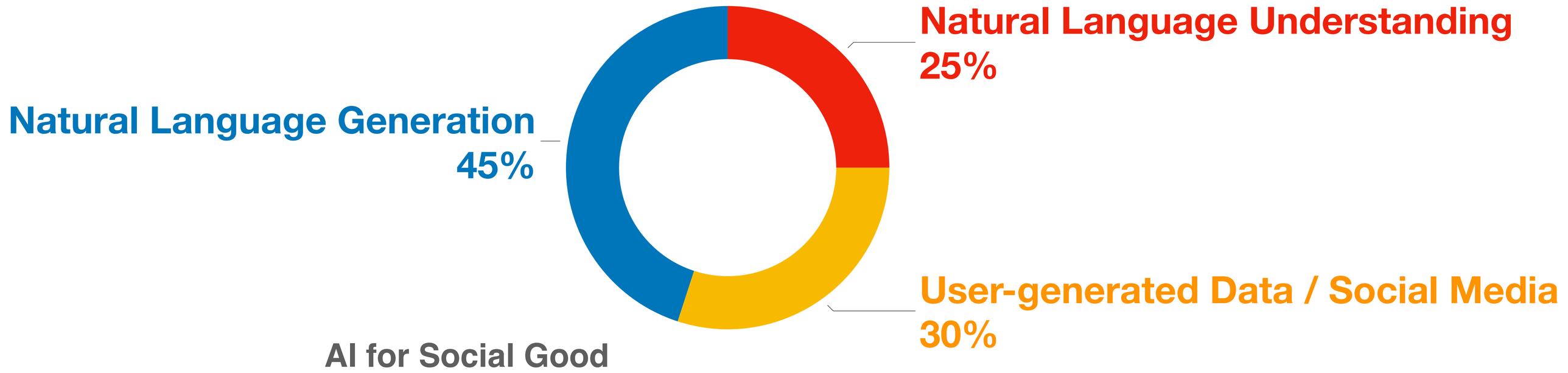
Natural Language Processing (AI group)

Wei Xu
School of Interactive Computing
Georgia Institute of Technology
@cocoweixu ✉ wei.xu@cc.gatech.edu
https://cocoxu.github.io/

My Research Group

Currently, 5 phd students + 5 undergraduate students + collaborators.



Natural Language Generation

Paraphrase Generation, Style Transfer, Text Simplification …

Slightly more fourth-graders nationwide are reading proficiently compared with a decade ago, but only a third of them are now reading well, according to a new report.

Most fourth-graders are better readers than they were 10 years ago.
But few of them can actually read well.

Natural Language Generation

Paraphrase Generation, Style Transfer, Text Simplification …

Slightly more fourth-graders ~~nationwide~~ are reading proficiently compared with a decade ago, but only a third of them are now reading well, ~~according to a new report.~~

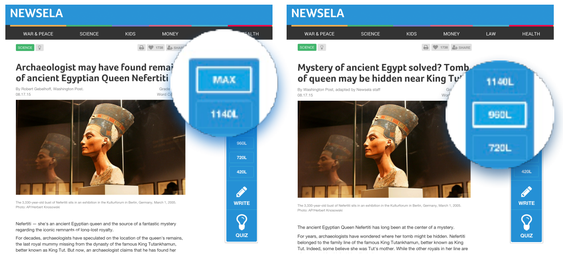
↓

Most fourth-graders are better readers than they were 10 years ago. But few of them can actually read well.

Natural Language Generation

This is now part of Google's TensorFlow library.

Newsela + Turk Corpus
(TACL 2015)



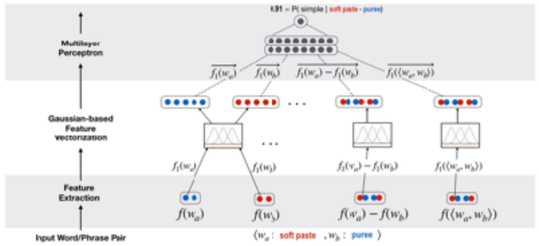
SARI Objective Function
(TACL 2016)

$$SARI = d_1 F_{add} + d_2 F_{keep} + d_3 P_{del}$$

$$p_{add}(n) = \frac{\sum_{g \in O} \min\left(\#_g(O \cap \overline{I}), \#_g(R)\right)}{\sum_{g \in O} \#_g(O \cap \overline{I})}$$

$$r_{add}(n) = \frac{\sum_{g \in O} \min\left(\#_g(O \cap \overline{I}), \#_g(R)\right)}{\sum_{g \in O} \#_g(R \cap \overline{I})}$$

Neural Readability Ranking
(EMNLP 2018)

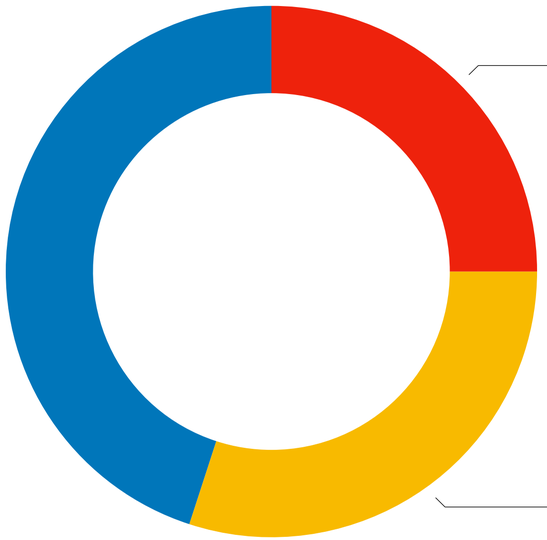


Semantic-based Alignment
(AAAI 2020; ACL 2020; and ongoing)

related to natural language understanding

Ongoing & Future Work: controllable generation, how to inject linguistic knowledge into neural networks, interactive learning models (NLP + HCI)…

Research Overview



Natural Language Understanding
25%

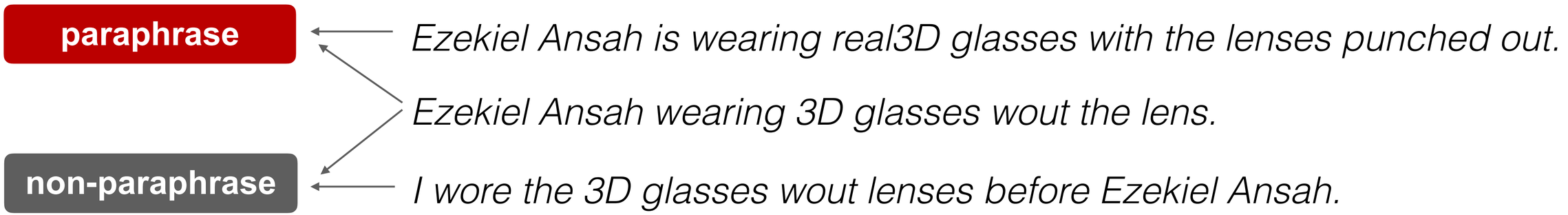
Natural Language Generation
45%

User-generated Data / Social Media
30%

AI for Social Good

Natural Language Understanding

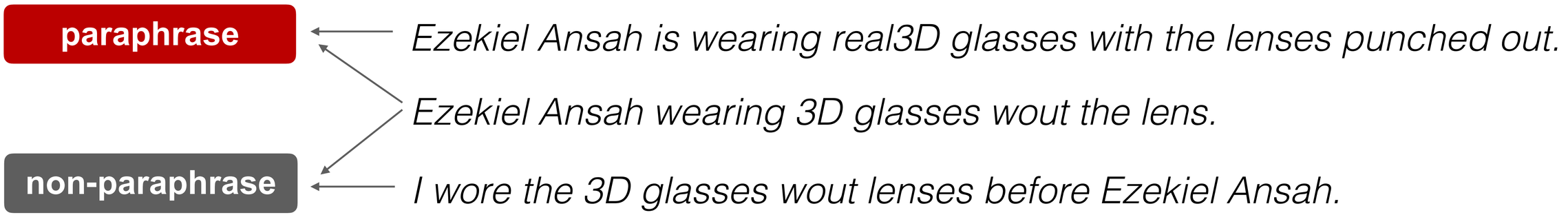
Paraphrase Identification



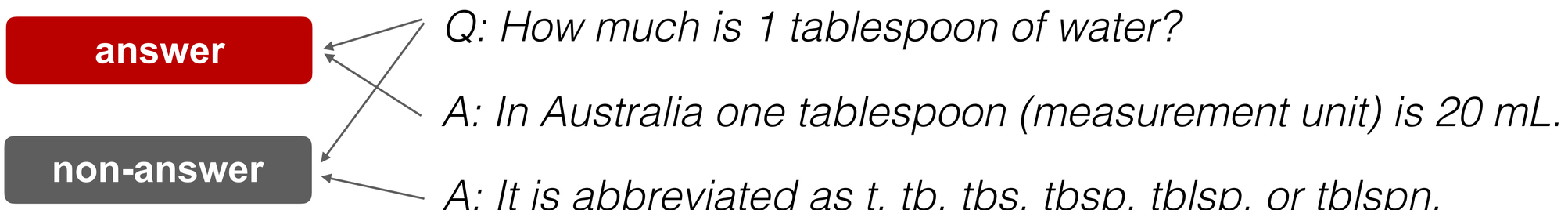
Wuwei Lan, Wei Xu. “Neural Network Models for Paraphrase Identification, Semantic Textual Similarity, Natural Language Inference, and Question Answering” Best Paper Award at COLING (2018)

Natural Language Understanding

Paraphrase Identification

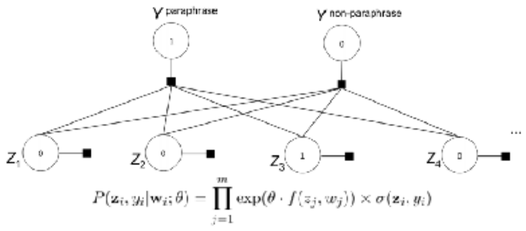


Question Answering



Semantic Similarity — our work

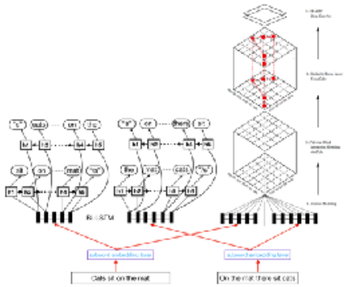
Multi-instance Learning
(TACL 2014)



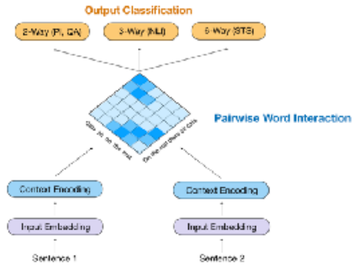
Twitter Paraphrase Corpus
(EMNLP 2017; ongoing)

related to natural language generation

Multi-task Subword Model
(NAACL 2018)



Pairwise Interaction Models
(COLING 2018; ongoing)



Semantic Similarity Models

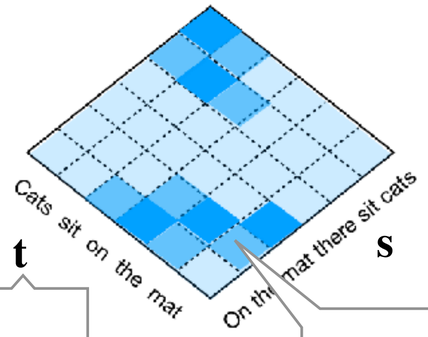
semi-Markov Conditional Random Fields for phrase alignment

$$\Psi(\mathbf{a}, \mathbf{s}, \mathbf{t}) = \sum_{i=1}^{|s|} score(i, a_i) + T(a_{i-l_i}, a_i)$$

$$P(\mathbf{a} \mid \mathbf{s}, \mathbf{t}) = \frac{exp\left(\Psi(\mathbf{a}, \mathbf{s}, \mathbf{t})\right)}{\sum_{\mathbf{a} \in \mathbf{A}} exp\left(\Psi(\mathbf{a}, \mathbf{s}, \mathbf{t})\right)}$$

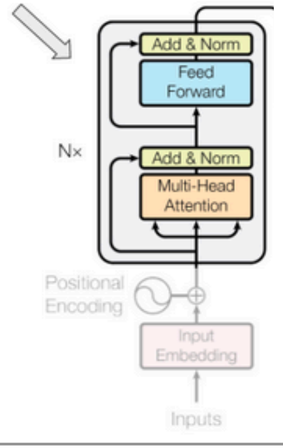
all possible alignments (dynamic programming)

Pairwise Phrase Interaction

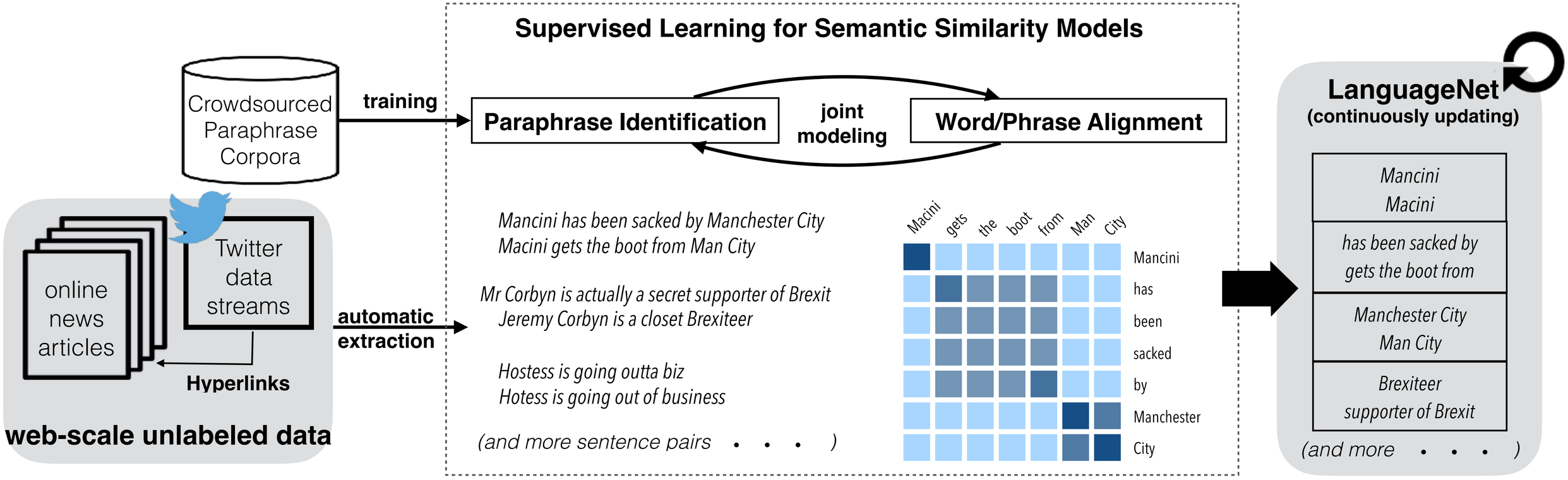


$$score(i,j) = d_h(i,j) + d_g(i,j) + d_b(i,j)$$

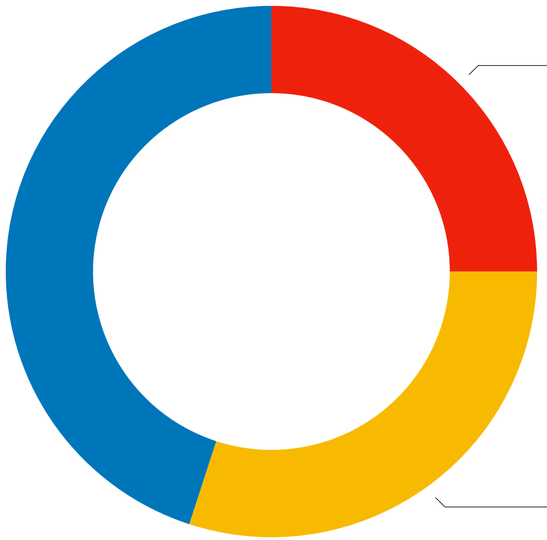
Transformer Encoder



LanguageNet 2.0 (ongoing & future work)



Research Overview



Natural Language Understanding
25%

Natural Language Generation
45%

User-generated Data / Social Media
30%

AI for Social Good

Extracting COVID-19 Events from Twitter

Slot	Text Span
Who	Her husband
Where	Columbus
Gender	Male
Age	39

Extracting COVID-19 Events from Twitter

- Models
  - Logistic Regression (LR)
  - BERT
- 65 train / 10 val / 25 test splits, evaluated by F1 score

		LR			BERT		
Slot	#	P	R	F1	P	R	F1
Who	450	.51	.48	.49	.76	.74	.75
C. contact	33	0	0	0	.33	.42	.37
Relation	11	0	0	0	.60	.55	.57
Employer	61	.32	.15	.20	.40	.39	.40
Recent v.	27	0	0	0	.50	.37	.43
Age	15	0	0	0	.56	.93	.70
Where	133	.27	.20	.23	.51	.59	.55
Gender M.	124	.31	.29	.30	.73	.60	.65
Gender F.	45	0	0	0	.71	.64	.67
When	14	0	0	0	.28	.69	.40

Joint work w/ Alan Ritter
(associate professor at IC @Georgia Tech)

How to get involved?

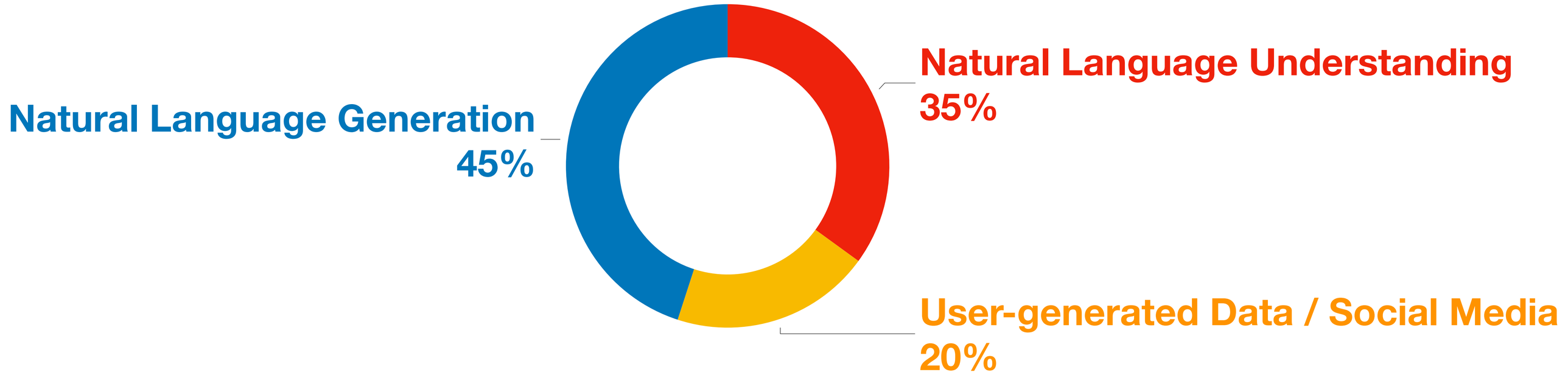
Some upcoming projects:

- Internet Slangs and Offensive Language
- Fake news / Misinformation
- User privacy on social media
- Lexical Simplification
- Interactive User Interface for Language Generation

Email me (wei.xu@cc.gatech.edu):

- Who you are? What you are good at?
- What you want to do? How big a project you want to take on?
- Why you want to do research? What is your career plan?

Outline


Natural Language Understanding
35%
Natural Language Generation
45%
User-generated Data / Social Media
20%