CONLL (3-4.11) & EMNLP (5-7.11) 2019

TL;DR

Tomasz Stanislawek

Applica seminar, 14.11.2019

Presentation plan

- Overview
- CONLL
- EMNLP keynotes
- EMNLP

CONLL/Workshops

Basic info

- 97 accepted papers
- Acceptance rate = 22.66%
- 2 papers from Poland
- 17 more workshops (3-4.11)
- 7 more tutorials (3-4.11)

CONLL/Workshops

Other interesting workshops

- DeepLo (Deep Learning for Low-Resource Natural Language Processing)
- MRQA (Machine Reading for Question Answering)
- AnnoNLP (Aggregating and analysing crowdsourced annotations for NLP)
- W-NUT (Workshop on Noisy User-generated Text)
- TextGraphs (Graph-Based Natural Language Processing)

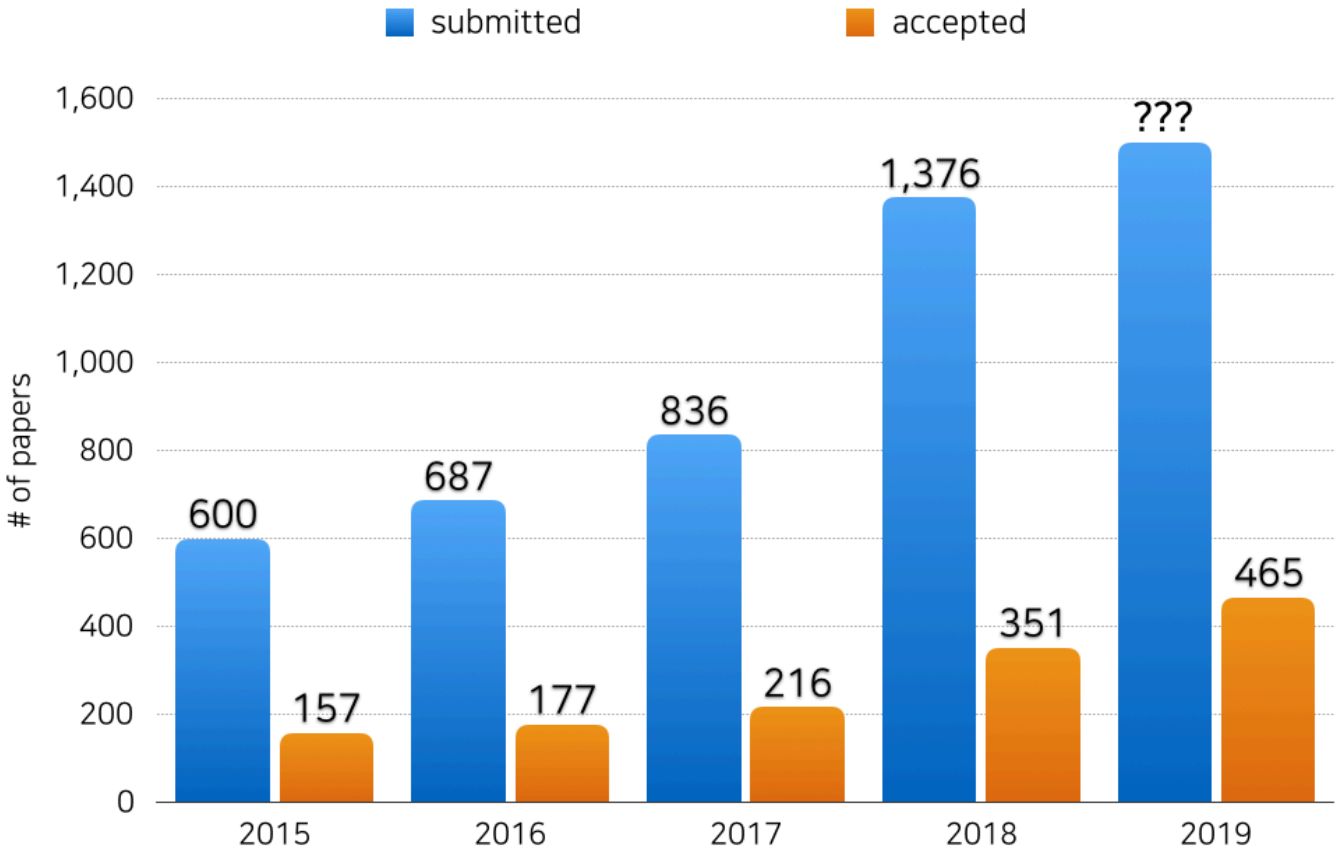
CONLL/Workshops

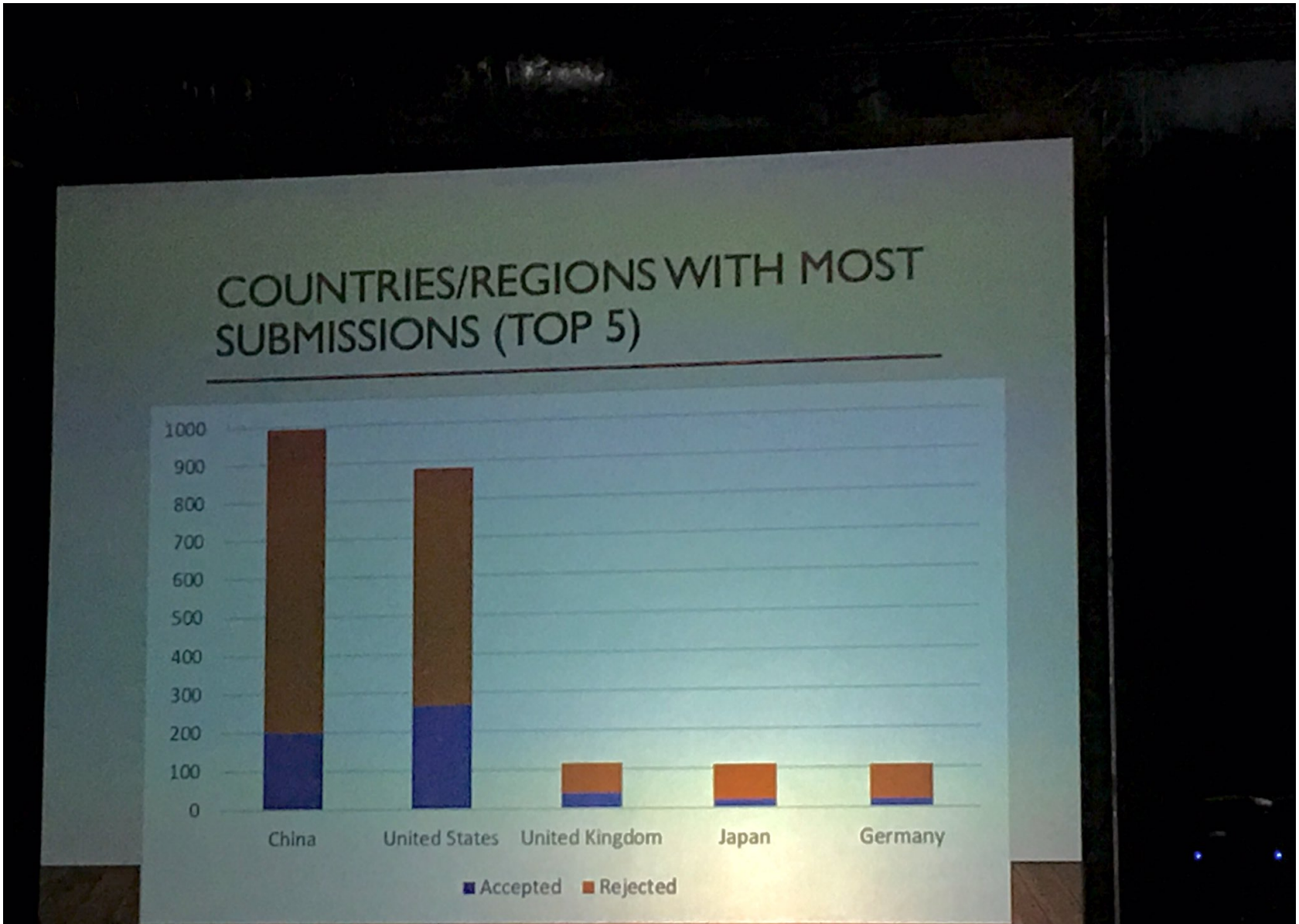
EMNLP

ACCEPTANCE STATISTICS

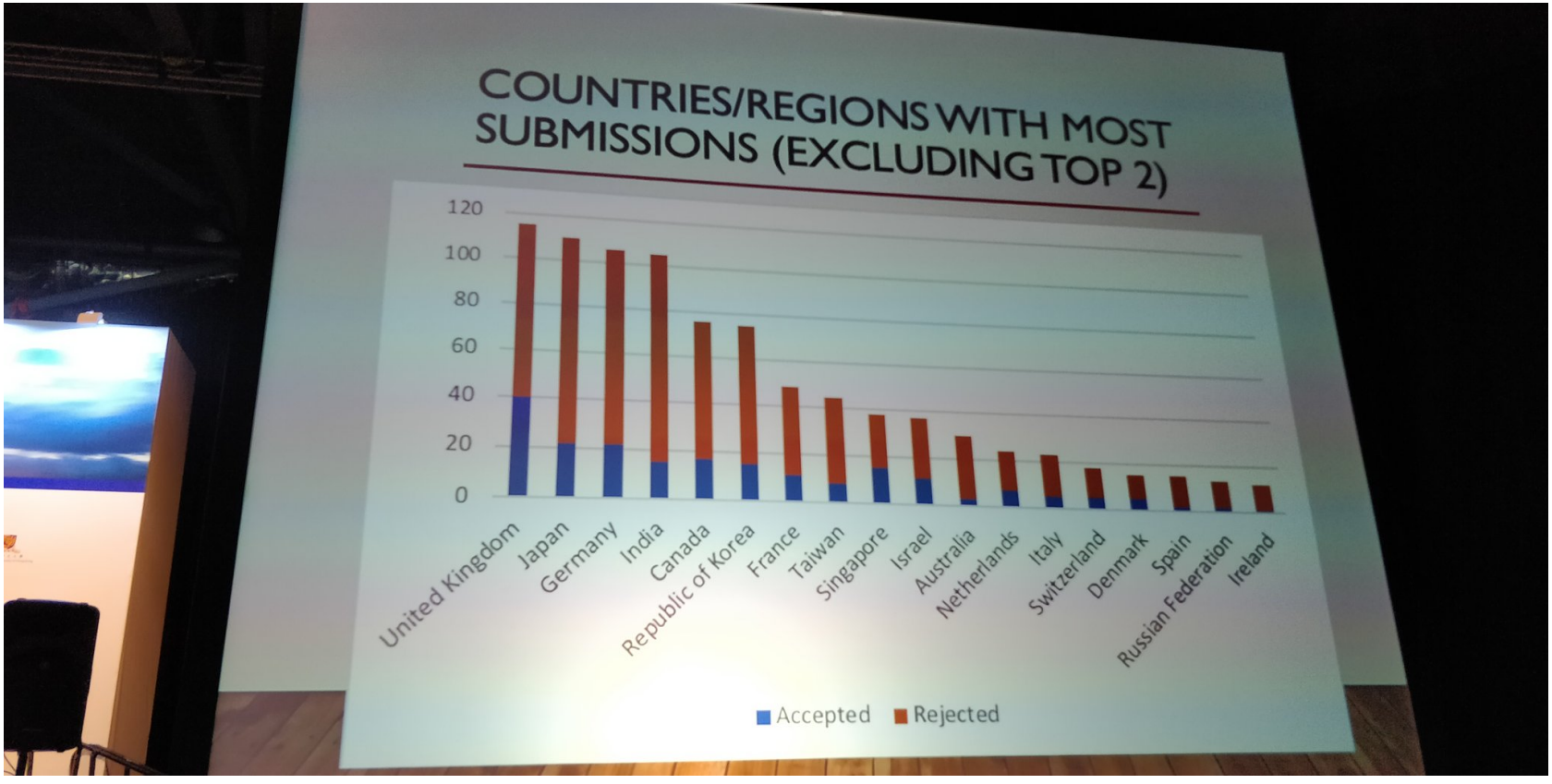
	Long Papers	Short Papers	Total
Reviewed	1813	1063	2876
Accepted as Oral	164	48	212
Accepted as Poster	301	170	471
Total Accepted	465 (25.6%)	218 (20.5%)	683 (23.7%)

EMNLP

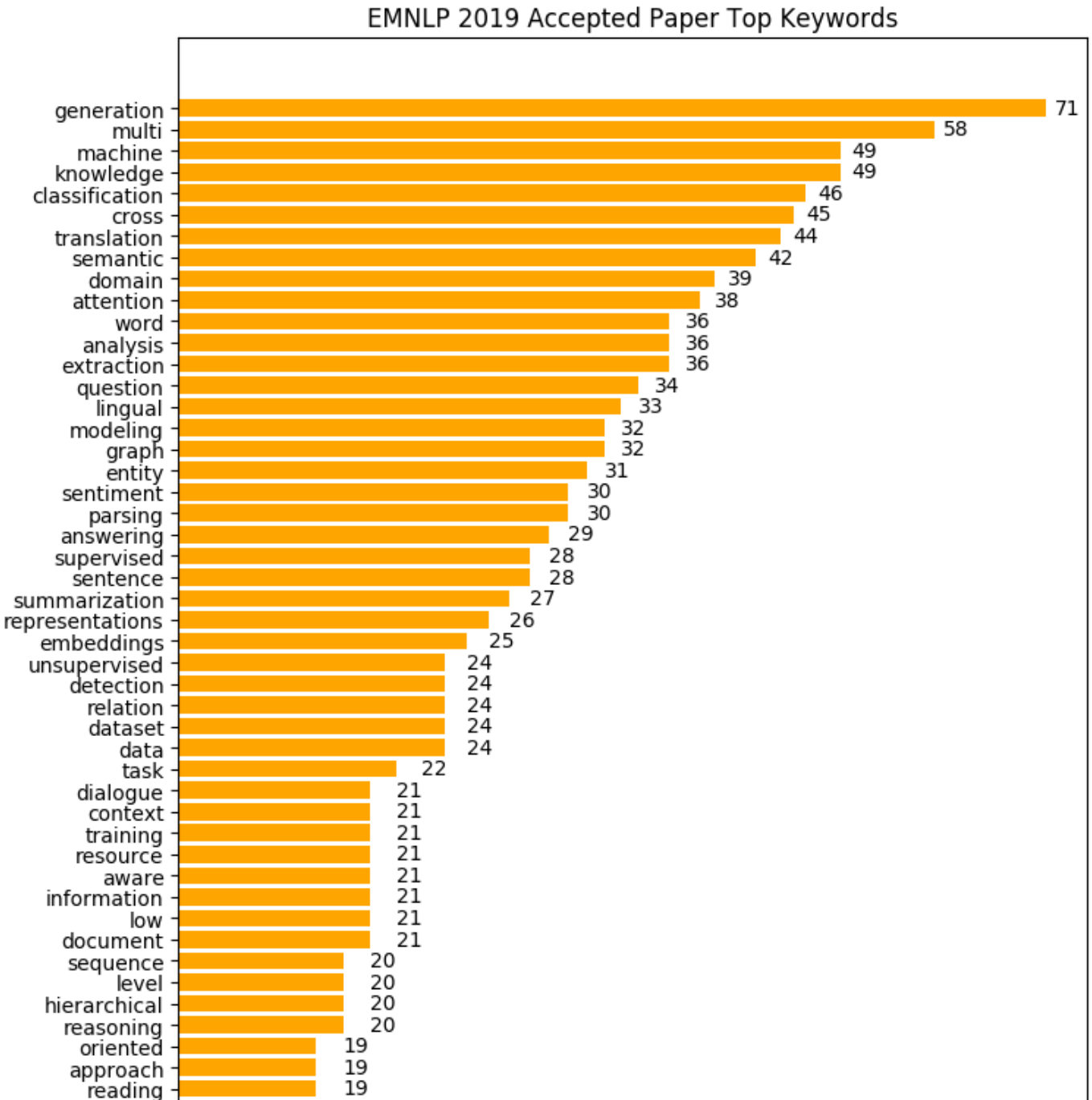

submitted
accepted
# of papers
1,600
1,400
1,200
1,000
800
600
400
200
0
600
157
687
177
836
216
1,376
351
???
465
2015
2016
2017
2018
2019


COUNTRIES/REGIONS WITH MOST SUBMISSIONS (TOP 5)
1000
900
800
700
600
500
400
300
200
100
0
China
United States
United Kingdom
Japan
Germany
Accepted
Rejected

EMNLP


COUNTRIES/REGIONS WITH MOST SUBMISSIONS (EXCLUDING TOP 2)
120
100
80
60
40
20
0
United Kingdom
Japan
Germany
India
Canada
Republic of Korea
France
Taiwan
Singapore
Israel
Australia
Netherlands
Italy
Switzerland
Denmark
Spain
Russian Federation
Ireland
Accepted
Rejected

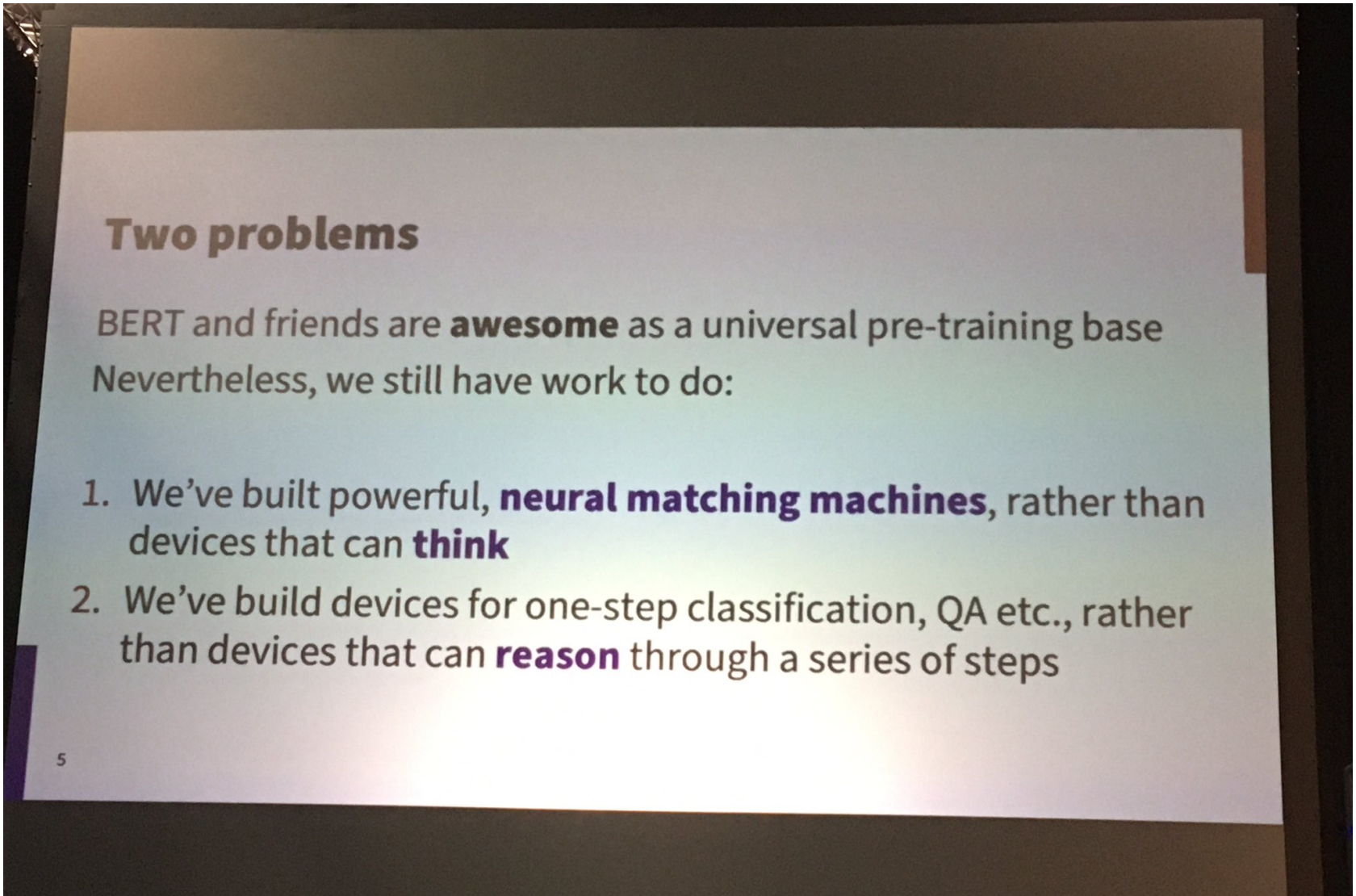
EMNLP


EMNLP 2019 Accepted Paper Top Keywords
generation 71
multi 58
machine 49
knowledge 49
classification 46
cross 45
translation 44
semantic 42
domain 39
attention 38
word 36
analysis 36
extraction 36
question 34
lingual 33
modeling 32
graph 32
entity 31
sentiment 30
parsing 30
answering 29
supervised 28
sentence 28
summarization 27
representations 26
embeddings 25
unsupervised 24
detection 24
relation 24
dataset 24
data 24
task 22
dialogue 21
context 21
training 21
resource 21
aware 21
information 21
low 21
document 21
sequence 20
level 20
hierarchical 20
reasoning 20
oriented 19
approach 19
reading 19

ACL 2019

	All submissions			Long submissions			Short submissions		
Country or Region	Sub.	Acc.	Rate (%)	Sub.	Acc.	Rate (%)	Sub.	Acc.	Rate (%)
Australia	46	11	23.9	22	4	18.2	24	7	29.2
Austria	5	0	0.0	2	0	0.0	3	0	0.0
Belgium	8	1	12.5	3	1	33.3	5	0	0.0
Brazil	11	0	0.0	6	0	0.0	5	0	0.0
Canada	74	16	21.6	44	12	27.3	30	4	13.3
Chile	2	0	0.0	2	0	0.0	0	0	N/A
China	817	155	19.0	567	118	20.8	250	37	14.8
Czech Republic	12	2	16.7	5	0	0.0	7	2	28.6
Denmark	25	4	16.0	11	1	9.1	14	3	21.4
Egypt	2	0	0.0	1	0	0.0	1	0	0.0
Estonia	2	0	0.0	2	0	0.0	0	0	N/A
Finland	6	0	0.0	2	0	0.0	4	0	0.0
France	60	11	18.3	32	4	12.5	28	7	25.0
Germany	136	39	28.7	73	26	35.6	63	13	20.6
Greece	7	4	57.1	1	1	100.0	6	3	50.0
Hong Kong	34	10	29.4	26	9	34.6	8	1	12.5
Hungary	7	1	14.3	3	1	33.3	4	0	0.0
India	107	18	16.8	54	16	29.6	53	2	3.8
Iran	3	0	0.0	2	0	0.0	1	0	0.0
Ireland	10	1	10.0	4	1	25.0	6	0	0.0
Israel	41	14	34.1	30	11	36.7	11	3	27.3
Italy	50	6	12.0	25	3	12.0	25	3	12.0
Japan	125	23	18.4	58	13	22.4	67	10	14.9
Luxembourg	2	0	0.0	2	0	0.0	0	0	N/A
Macau	5	1	20.0	3	1	33.3	2	0	0.0
Malta	2	0	0.0	0	0	N/A	2	0	0.0
Mexico	2	0	0.0	0	0	N/A	2	0	0.0
Netherlands	36	9	25.0	22	8	36.4	14	1	7.1
Norway	6	2	33.3	4	1	25.0	2	1	50.0
Pakistan	2	0	0.0	1	0	0.0	1	0	0.0
Peru	2	0	0.0	1	0	0.0	1	0	0.0
Poland	7	1	14.3	5	1	20.0	2	0	0.0
Portugal	8	3	37.5	4	2	50.0	4	1	25.0
Qatar	4	0	0.0	2	0	0.0	2	0	0.0
Republic of Korea	72	7	9.7	36	4	11.1	36	3	8.3
Romania	2	1	50.0	2	1	50.0	0	0	N/A
Russian Federation	14	4	28.6	7	2	28.6	7	2	28.6
Singapore	46	16	34.8	39	13	33.3	7	3	42.9
Slovakia	2	0	0.0	1	0	0.0	1	0	0.0
South Africa	2	1	50.0	1	0	0.0	1	1	100
Spain	29	6	20.7	12	1	8.3	17	5	29.4
Sri Lanka	5	0	0.0	1	0	0.0	4	0	0.0
Sweden	9	0	0.0	4	0	0.0	5	0	0.0
Switzerland	23	4	17.4	12	2	16.7	11	2	18.2
Taiwan	46	6	13.0	18	3	16.7	28	3	10.7
Thailand	2	0	0.0	1	0	0.0	1	0	0.0
Turkey	7	0	0.0	3	0	0.0	4	0	0.0
United Arab Emirates	4	2	50.0	1	1	100.0	3	1	33.3
United Kingdom	138	41	29.7	84	30	35.7	54	11	20.4
United States	820	236	28.8	485	154	31.8	335	82	24.5
Others	18	2		12	0		6	3	
TOTAL	**2905**	**660**	22.7	**1737**	**447**	25.7	**1168**	**213**	18.2

CONLL: C. Manning - Multi-step reasoning for answering complex questions



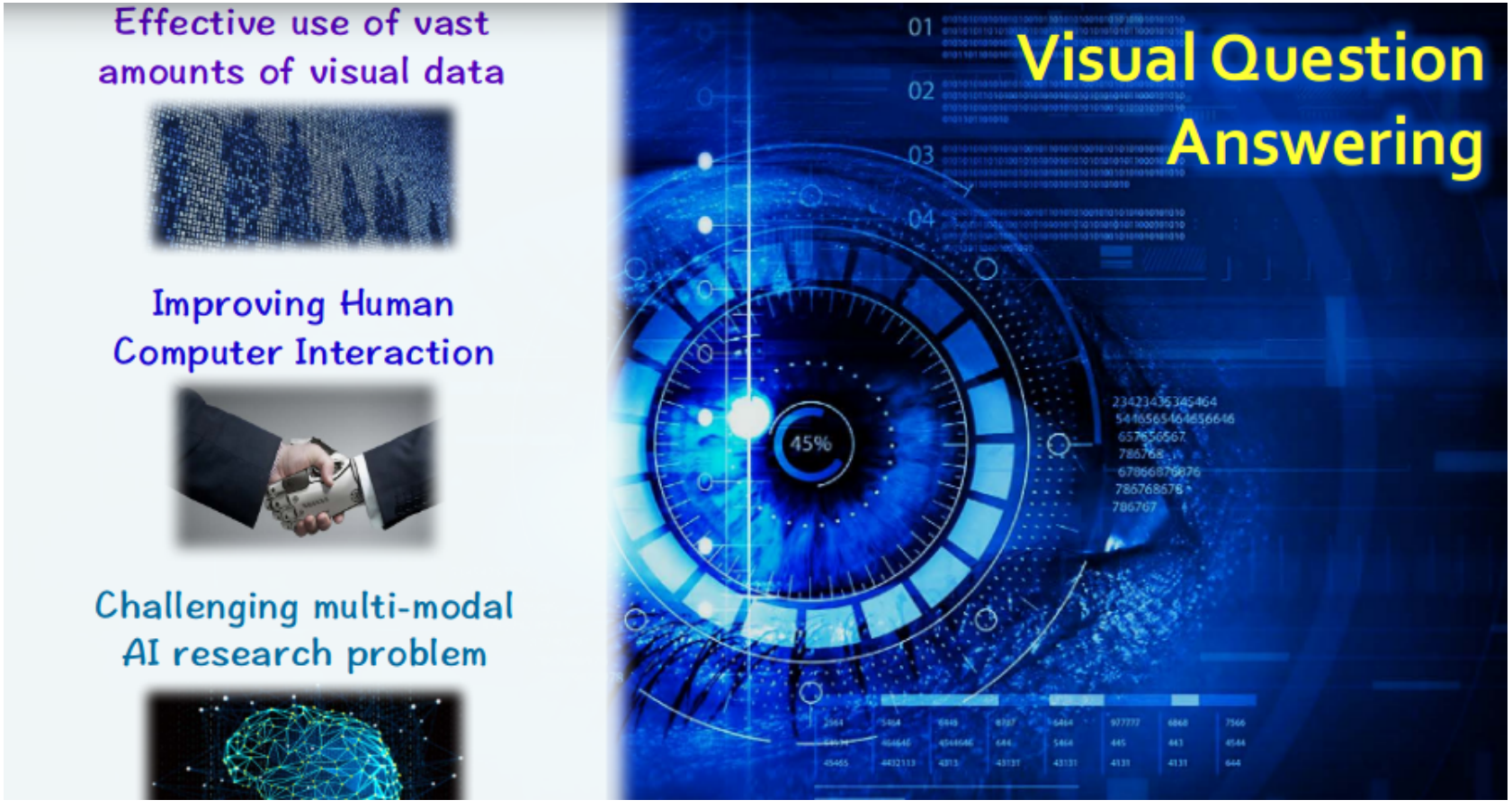
CONLL: C. Manning - Multi-step reasoning for answering complex questions

https://cs.stanford.edu/people/dorarad/gqa/slides.html

"When a person understands a story, [they] can demonstrate [their] understanding by answering questions about the story. Since questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding."

— Wendy Lehnert (PhD, 1977)

CONLL: C. Manning - Multi-step reasoning for answering complex questions



CONLL: C. Manning - Multi-step reasoning for answering complex questions

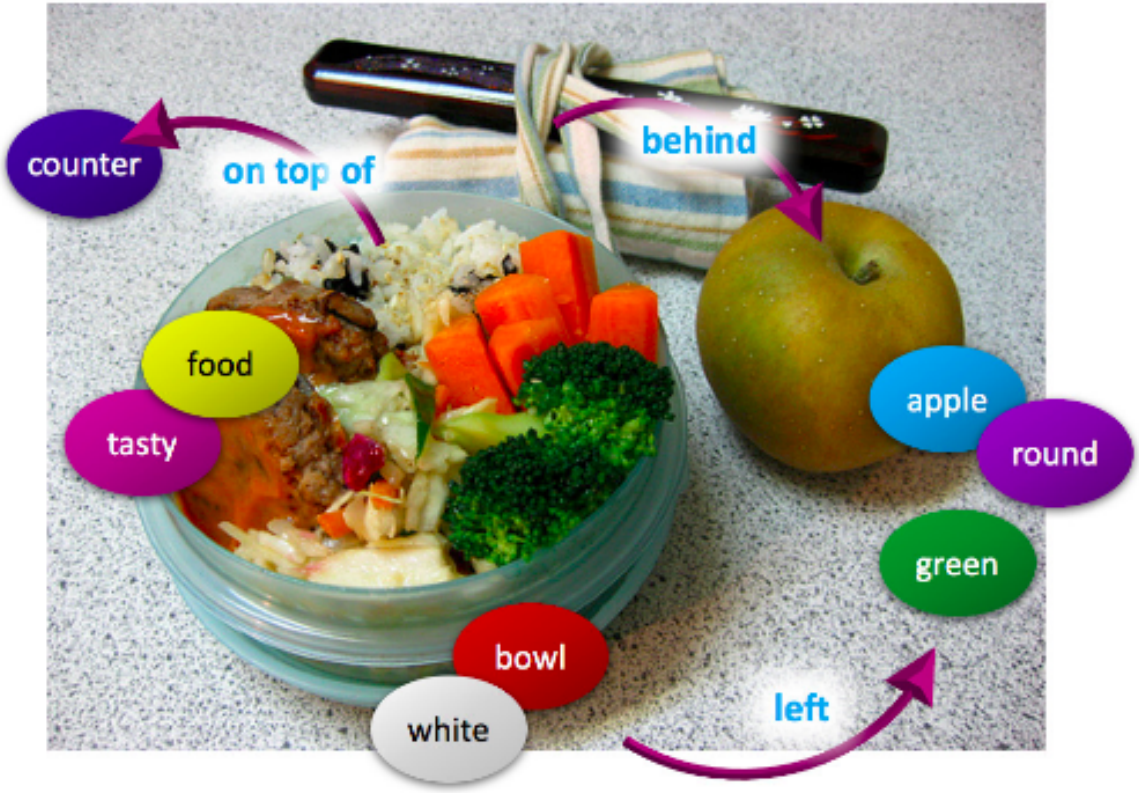


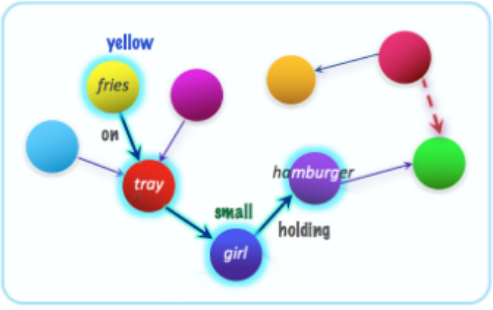
Figure 1: Examples from the new GQA dataset for visual reasoning and compositional question answering:
Is the bowl to the right of the green apple?
What type of fruit in the image is round?
What color is the fruit on the right side, red or green?
Is there any milk in the bowl to the left of the apple?

CONLL: C. Manning - Multi-step reasoning for answering complex questions

Pattern: What|Which <type> [do you think] <is> <dobject>, <attr> or <decoy>?
Program: **Select:** <dobject> → **Choose** <type>: <attr>|<decoy>
Reference: The food on the red object left of the small girl that is holding a hamburger
Decoy: brown

What color is the food on the red object left of the small girl that is holding a hamburger, yellow or brown?

Select: hamburger → **Relate:** girl, holding → **Filter** size: small → **Relate:** object, left → **Filter** color: red → **Relate:** food, on → **Choose** color: yellow | brown



Graph Normalization

- Ontology construction
- Edge Pruning
- Object Augmentation
- Global Properties

Question Generation

- Patterns Collection
- Compositional References
- Decoys Selection
- Probabilistic Generation

Sampling and Balancing

- Distribution Balancing
- Type-Based Sampling
- Deduplication

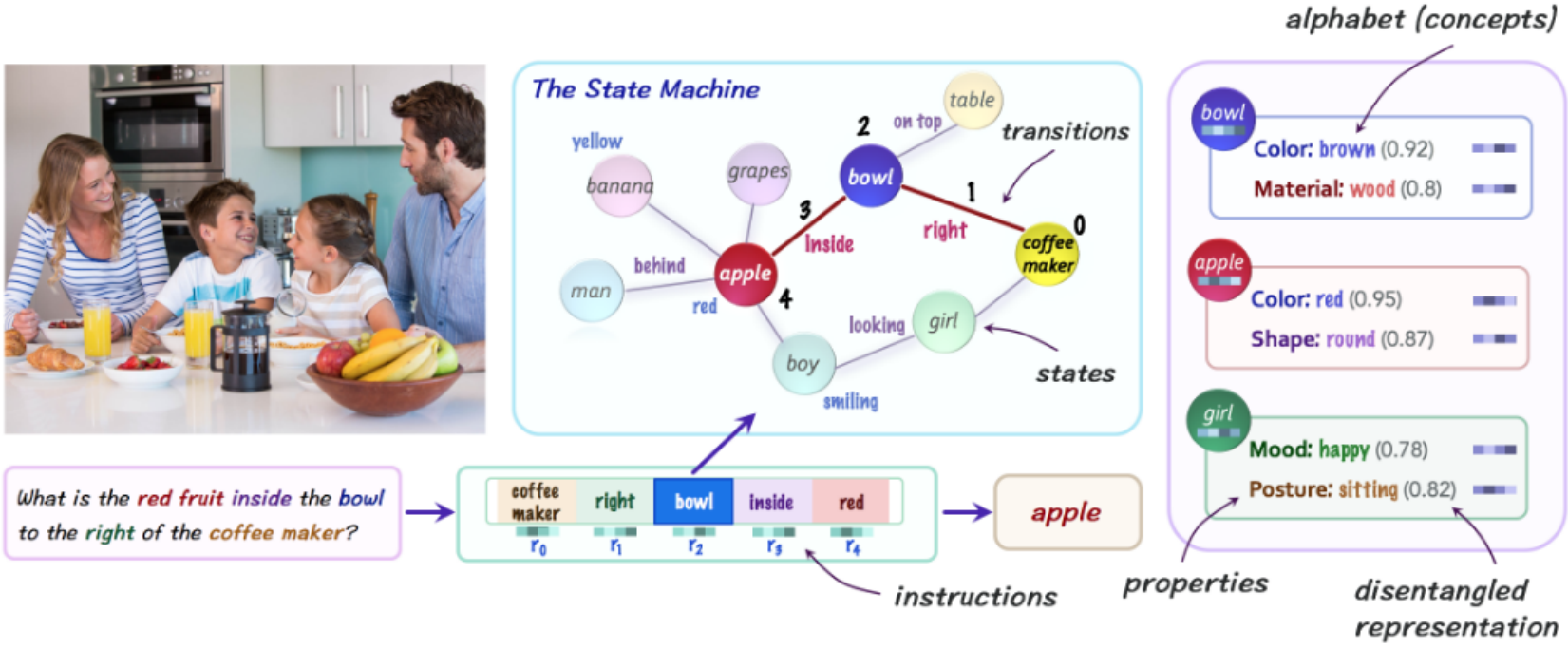
Entailments Relations

- Functional Programs
- Entailment Relations
- Recursive Reachability

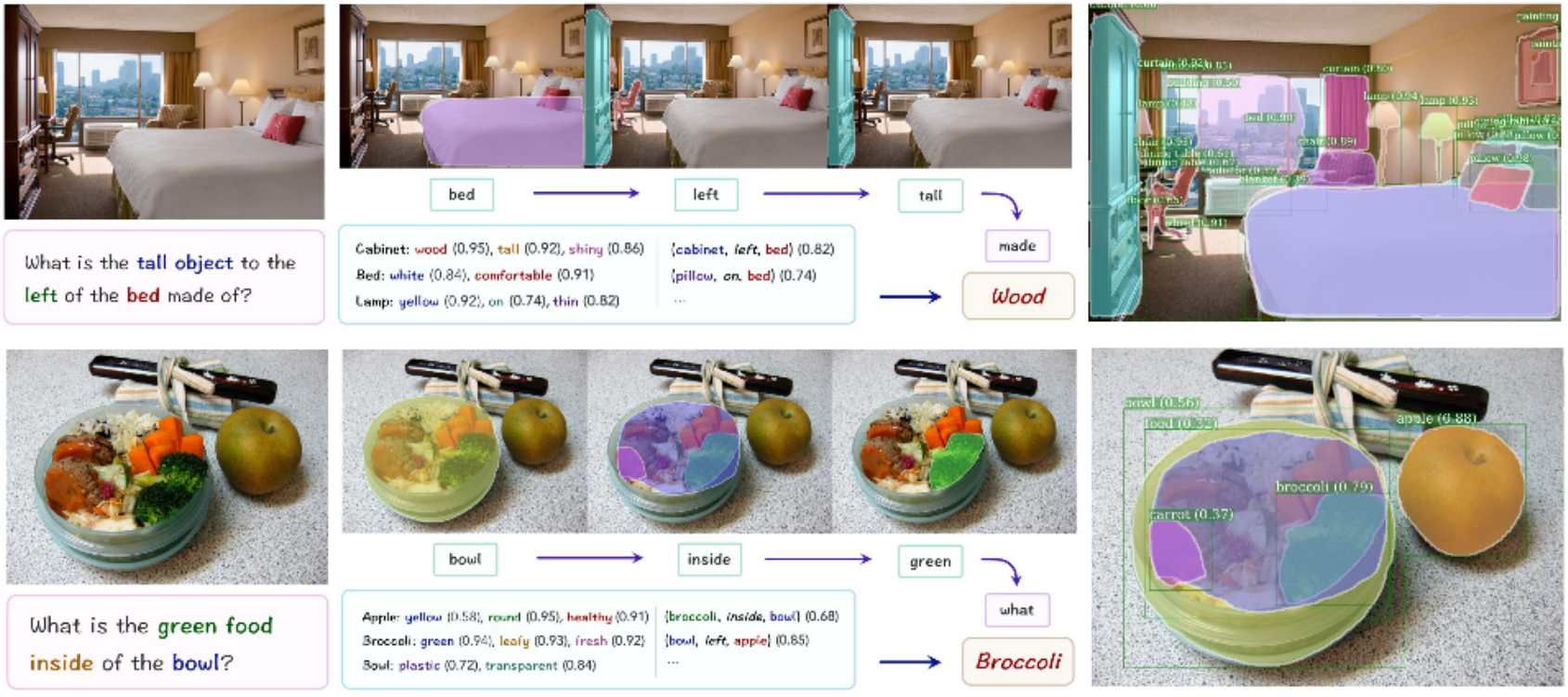
New Metrics

- Consistency
- Validity & Plausibility
- Distribution
- Grounding

CONLL: C. Manning - Multi-step reasoning for answering complex questions



CONLL: C. Manning - Multi-step reasoning for answering complex questions



CONLL: C. Manning - Multi-step reasoning for answering complex questions

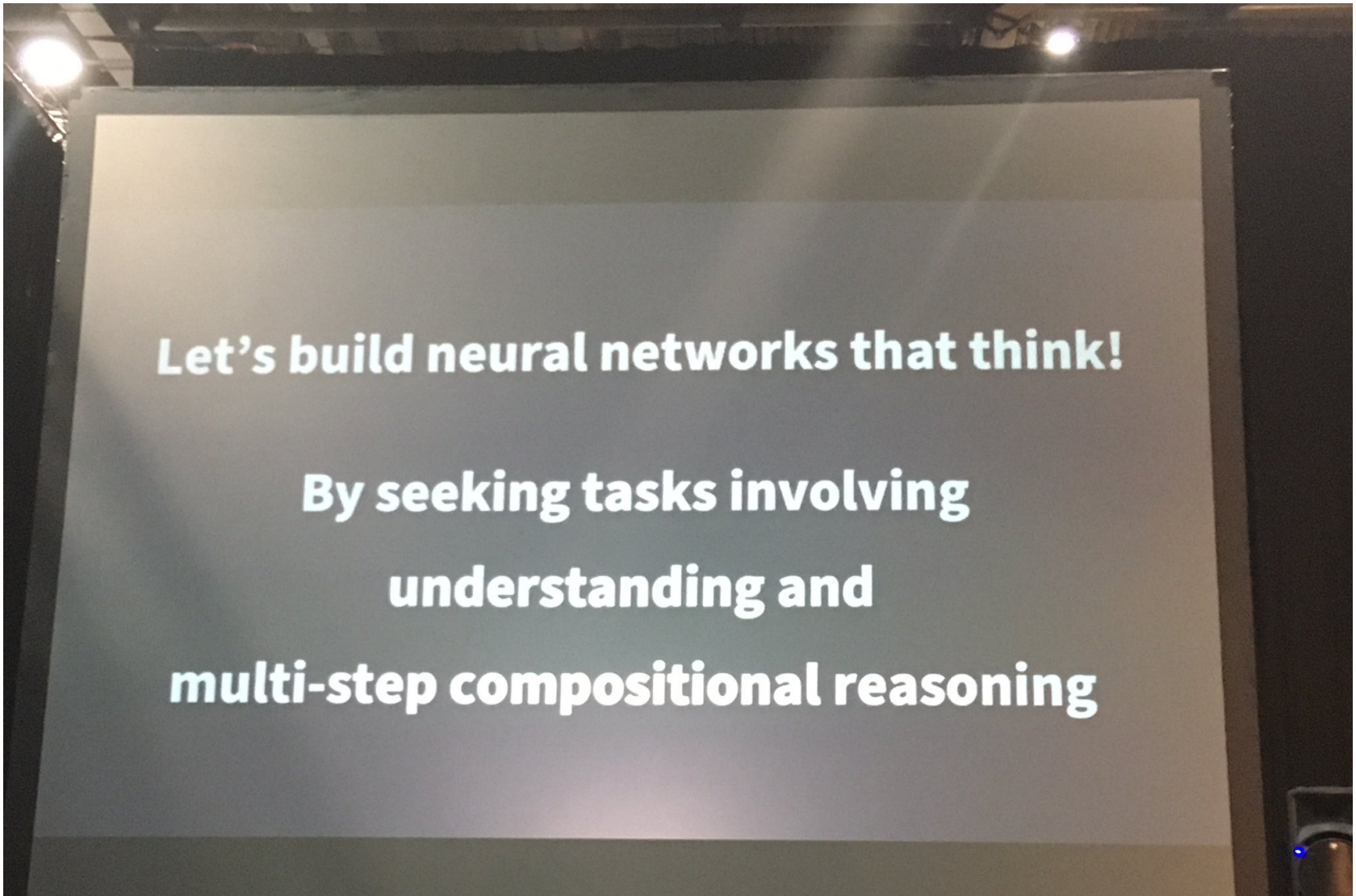
Table 2: GQA ensemble

Model	Accuracy
Kakao*	**73.33**
270	70.23
NSM	67.25
LXRT	62.71
GRN	61.22
MSM	61.09
DREAM	60.93
SK T-Brain*	60.87
PKU	60.79
Musan	59.93

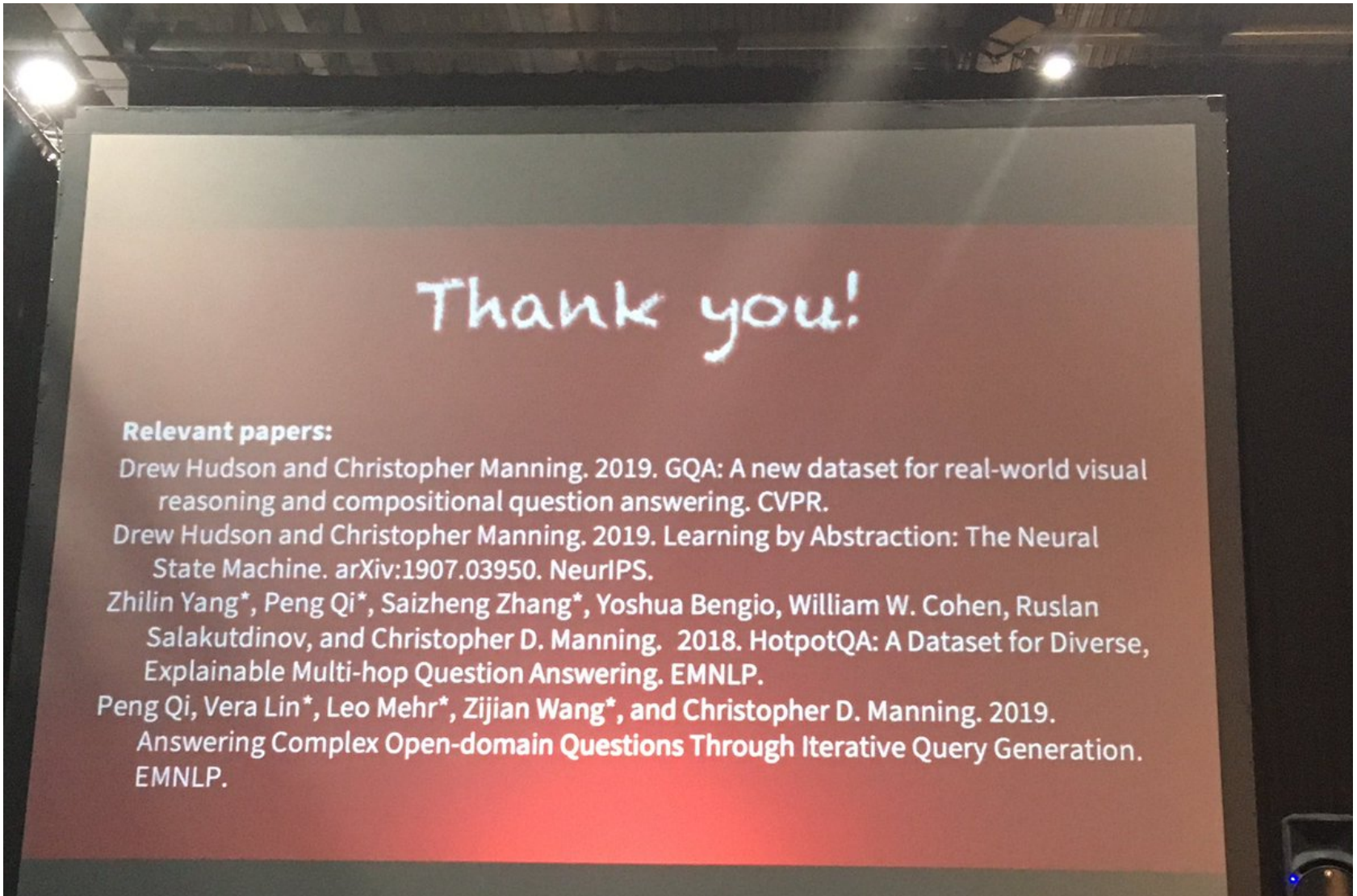
Table 3: VQA-CPv2

Model	Accuracy
SAN [83]	24.96
HAN [57]	28.65
GVQA [3]	31.30
RAMEN [70]	39.21
BAN [44]	39.31
MuRel [15]	39.54
ReGAT [50]	40.42
NSM	**45.80**

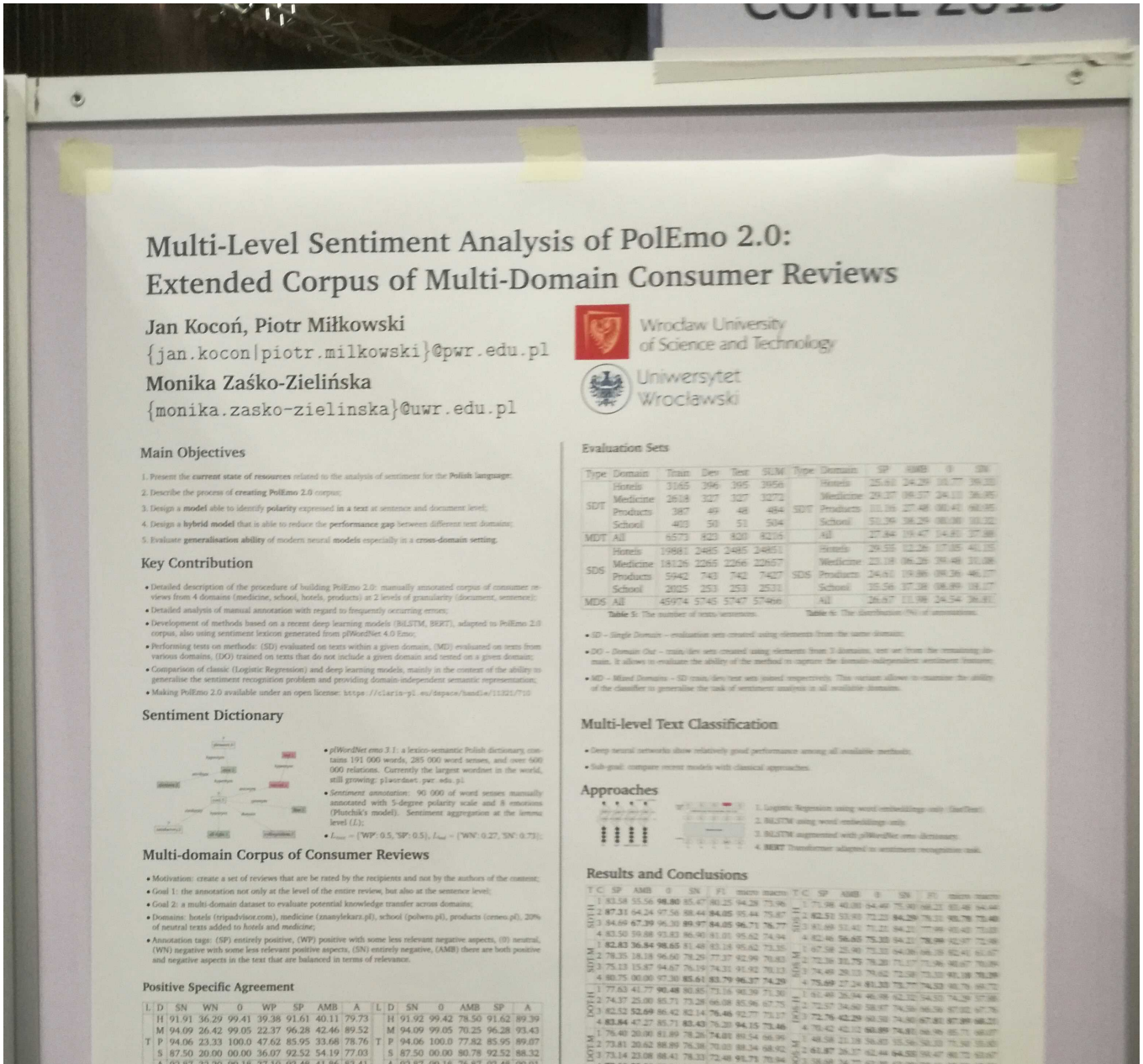
CONLL: C. Manning - Multi-step reasoning for answering complex questions



CONLL: C. Manning - Multi-step reasoning for answering complex questions



CONLL: J. Kocoń - PolEmo 2.0 Sentiment Analysis Dataset for CoNLL



Multi-Level Sentiment Analysis of PolEmo 2.0: Extended Corpus of Multi-Domain Consumer Reviews

Jan Kocoń, Piotr Miłkowski
{jan.kocon|piotr.milkowski}@pwr.edu.pl
Wrocław University of Science and Technology

Monika Zaśko-Zielińska
{monika.zasko-zielinska}@uwr.edu.pl
Uniwersytet Wrocławski

Main Objectives

1. Present the **current state of resources** related to the analysis of sentiment for the **Polish language**;
2. Describe the process of **creating PolEmo 2.0** corpus;
3. Design a **model** able to identify **polarity** expressed **in a text** at sentence and document level;
4. Design a **hybrid model** that is able to reduce the **performance gap** between different test domains;
5. Evaluate **generalisation ability** of modern neural **models** especially in a **cross-domain setting**.

Key Contribution

- Detailed description of the procedure of building PolEmo 2.0: manually annotated corpus of consumer reviews from 4 domains (medicine, school, hotels, products) at 2 levels of granularity (document, sentence);
- Detailed analysis of manual annotation with regard to frequently occurring errors;
- Development of methods based on a recent deep learning models (BiLSTM, BERT), adapted to PolEmo 2.0 corpus, also using sentiment lexicon generated from plWordNet 4.0 Emo;
- Performing tests on methods: (SD) evaluated on texts within a given domain, (MD) evaluated on texts from various domains, (DO) trained on texts that do not include a given domain and tested on a given domain;
- Comparison of classic (Logistic Regression) and deep learning models, mainly in the context of the ability to generalise the sentiment recognition problem and providing domain-independent semantic representation;
- Making PolEmo 2.0 available under an open license: https://clarin-pl.eu/dspace/handle/11321/710

Sentiment Dictionary

- *plWordNet emo 3.1*: a lexico-semantic Polish dictionary, contains 191 000 words, 285 000 word senses, and over 600 000 relations. Currently the largest wordnet in the world, still growing: plwordnet.pwr.edu.pl.
- *Sentiment annotation*: 90 000 of word senses manually annotated with 5-degree polarity scale and 8 emotions (Plutchik's model). Sentiment aggregation at the *lexeme* level (L):
- $L_{sum} = \{'WP': 0.5, 'SP': 0.5\}$, $L_{sat} = \{'WN': 0.27, 'SN': 0.73\}$;

Multi-domain Corpus of Consumer Reviews

- Motivation: create a set of reviews that are be rated by the recipients and not by the authors of the content;
- Goal 1: the annotation not only at the level of the entire review, but also at the sentence level;
- Goal 2: a multi-domain dataset to evaluate potential knowledge transfer across domains;
- Domains: hotels (ripadvisor.com), medicine (znanylekarz.pl), school (polwro.pl), products (ceneo.pl), 20% of neutral texts added to *hotels* and *medicine*;
- Annotation tags: (SP) entirely positive, (WP) positive with some less relevant negative aspects, (0) neutral, (WN) negative with some less relevant positive aspects, (SN) entirely negative, (AMB) there are both positive and negative aspects in the text that are balanced in terms of relevance.

Positive Specific Agreement

L	D	SN	WN	0	WP	SP	AMB	A
	H	91.91	36.29	99.41	39.38	91.61	40.11	79.73
	M	94.09	26.42	99.05	22.37	96.28	42.46	89.52
T	P	94.06	23.33	100.0	47.62	85.95	33.68	78.76
	S	87.50	20.00	00.00	36.07	92.52	54.19	77.03

L	D	SN	0	AMB	SP	A
	H	91.92	99.42	78.50	91.62	89.39
	M	94.09	99.05	70.25	96.28	93.43
T	P	94.06	100.0	77.82	85.95	89.07
	S	87.50	00.00	80.78	92.52	88.32

Evaluation Sets

Type	Domain	Train	Dev	Test	SUM
SDT	Hotels	3165	396	395	3956
	Medicine	2618	327	327	3272
	Products	387	49	48	484
	School	403	50	51	504
MDT	All	6573	823	820	8216
SDS	Hotels	19881	2485	2485	24851
	Medicine	18126	2265	2266	22657
	Products	5942	743	742	7427
	School	2025	253	253	2531
MDS	All	45974	5745	5747	57466

Table 5: The number of texts/sentences.

Type	Domain	SP	AMB	0	SN
SDT	Hotels	25.61	24.29	10.77	39.33
	Medicine	29.37	08.57	24.11	36.95
	Products	11.16	27.48	08.47	52.89
	School	31.39	38.29	08.08	22.22
	All	27.84	19.47	14.85	37.88
SDS	Hotels	29.55	12.26	17.05	41.15
	Medicine	23.19	06.29	39.48	31.08
	Products	24.61	19.86	09.36	46.17
	School	35.56	37.28	08.89	18.17
	All	26.67	11.98	24.54	36.81

Table 6: The distribution [%] of annotations.

- *SD – Single Domain* – evaluation sets created using elements from the same domain;
- *DO – Domain Out* – train/dev sets created using elements from 3 domains, test set from the remaining domain. It allows to evaluate the ability of the method to capture the domain-independent sentiment features;
- *MD – Mixed Domains* – SD train/dev/test sets joined respectively. This variant allows to examine the ability of the classifier to generalise the task of sentiment analysis in all available domains.

Multi-level Text Classification

- Deep neural networks show relatively good performance among all available methods;
- Sub-goal: compare recent models with classical approaches.

Approaches

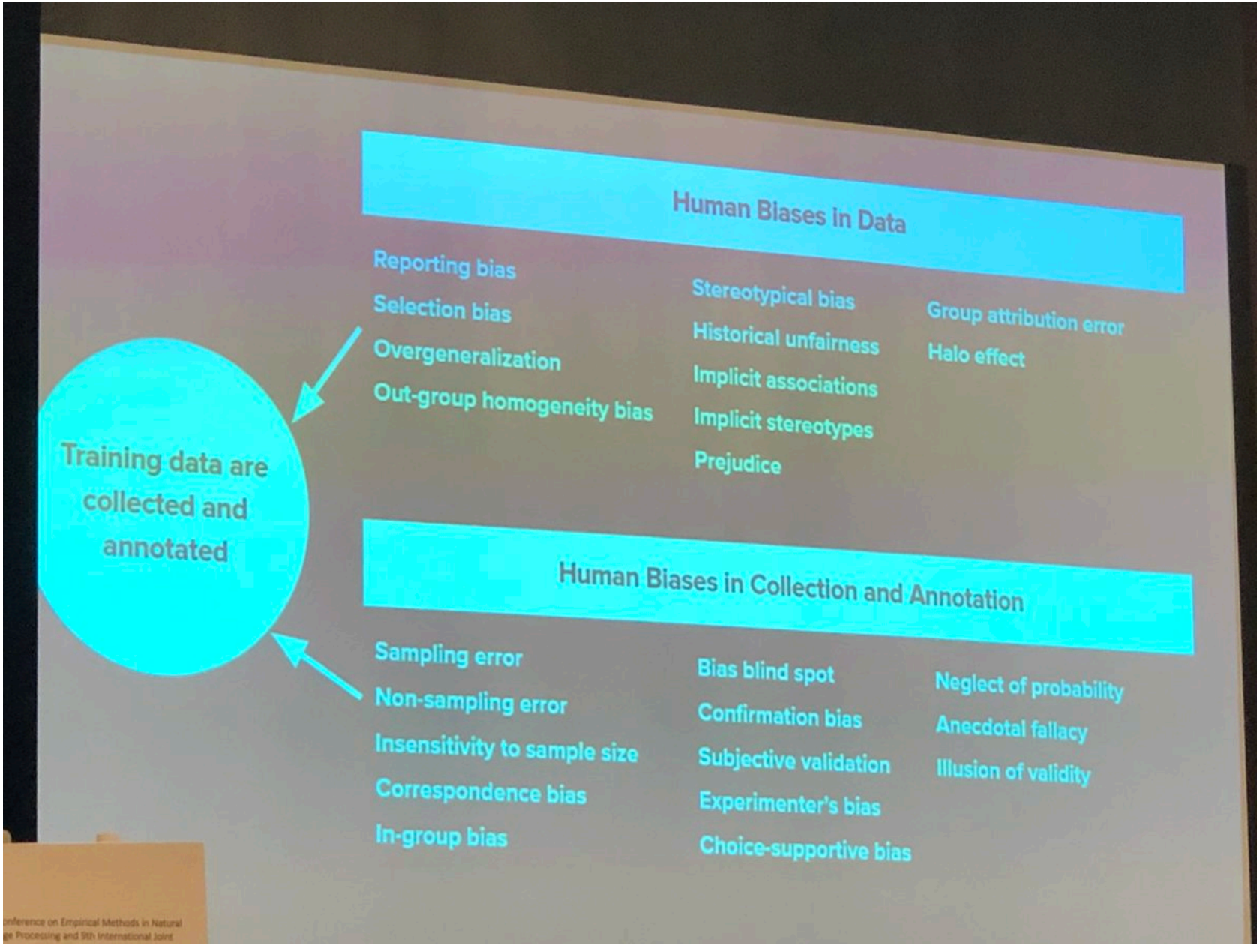
1. Logistic Regression using word embeddings only (fastText);
2. BiLSTM using word embeddings only;
3. BiLSTM augmented with plWordNet emo dictionary;
4. BERT Transformer adapted to sentiment recognition task.

Results and Conclusions

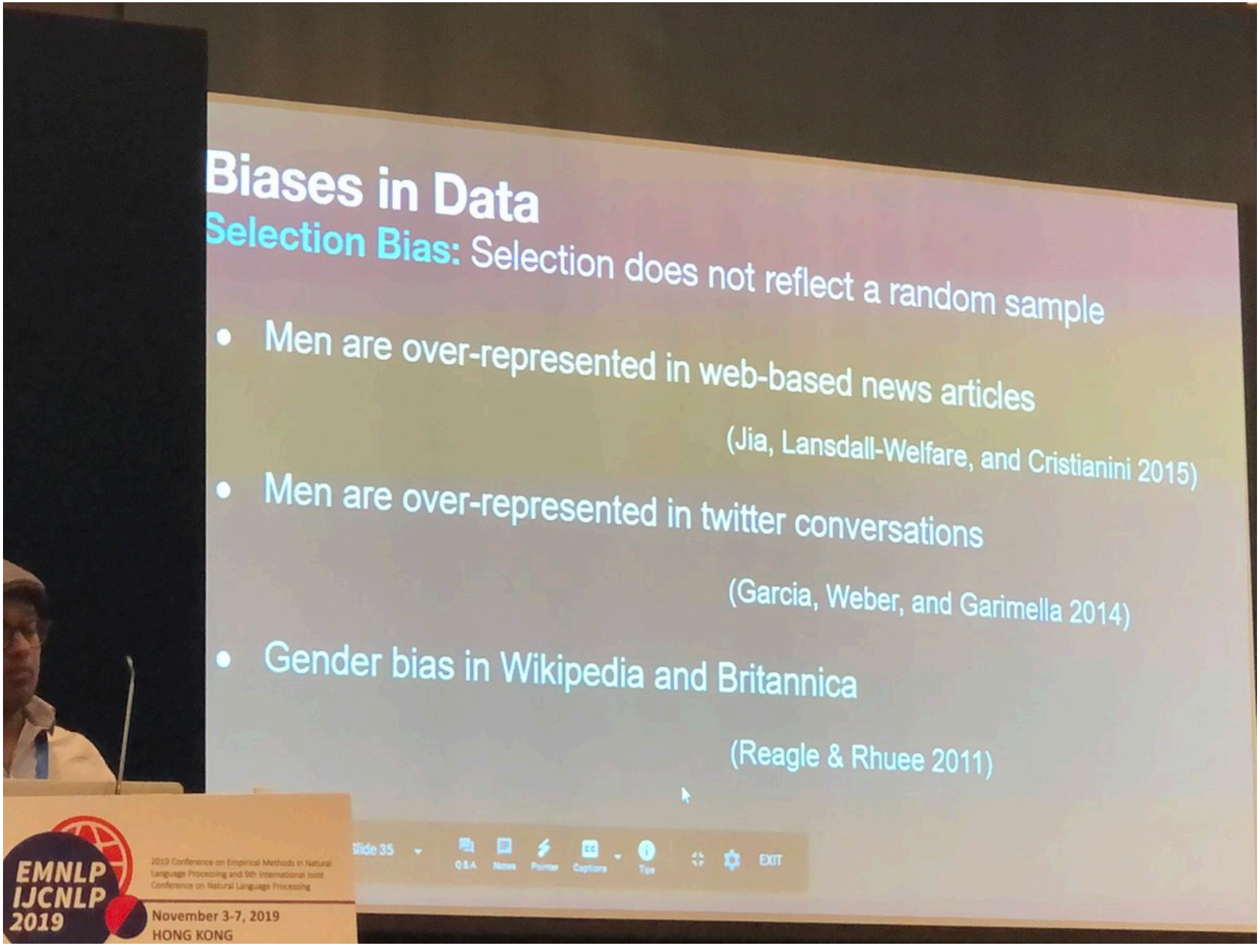
CONLL: J. Kocoń - PolEmo 2.0 Sentiment Analysis Dataset for CoNLL

https://sentimenti.pl/
In this article we present an extended version of PolEmo – a corpus of consumer reviews from 4 domains: medicine, hotels, products and school. Current version (PolEmo 2.0) contains 8,216 reviews having 57,466 sentences. Each text and sentence was manually annotated with sentiment in 2+1 scheme, which gives a total of 197,046 annotations. We obtained a high value of Positive Specific Agreement, which is 0.91 for texts and 0.88 for sentences. PolEmo 2.0 is publicly available under a Creative Commons copyright license.We explored recent deep learning approaches for the recognition of sentiment, such as Bi-directional Long Short-Term Memory (BiL-STM) and Bidirectional Encoder Representations from Transformers (BERT).

CONLL: Biases in Data/ML/NLP



CONLL: Biases in Data/ML/NLP



CONLL: Biases in Data/ML/NLP

http://www.datascienceassn.org/content/cognitive-bias-data-science

20 COGNITIVE BIASES THAT SCREW UP YOUR DECISIONS

1. Anchoring bias.

People are **over-reliant** on the first piece of information they hear. In a salary negotiation, whoever makes the first offer establishes a range of reasonable possibilities in each person's mind.

2. Availability heuristic.

People **overestimate the importance** of information that is available to them. A person might argue that smoking is not unhealthy because they know someone who lived to 100 and smoked three packs a day.

3. Bandwagon effect.

The probability of one person adopting a belief increases based on the number of people who hold that belief. This is a powerful form of **groupthink** and is reason why meetings are often unproductive.

4. Blind-spot bias.

Failing to recognize your own cognitive biases is a bias in itself. People notice cognitive and motivational biases much more in others than in themselves.

5. Choice-supportive bias.

When you choose something, you tend to feel positive about it, even if that **choice has flaws**. Like how you think your dog is awesome — even if it bites people every once in a while.

6. Clustering illusion.

This is the tendency to **see patterns in random events**. It is key to various gambling fallacies, like the idea that red is more or less likely to turn up on a roulette table after a string of reds.

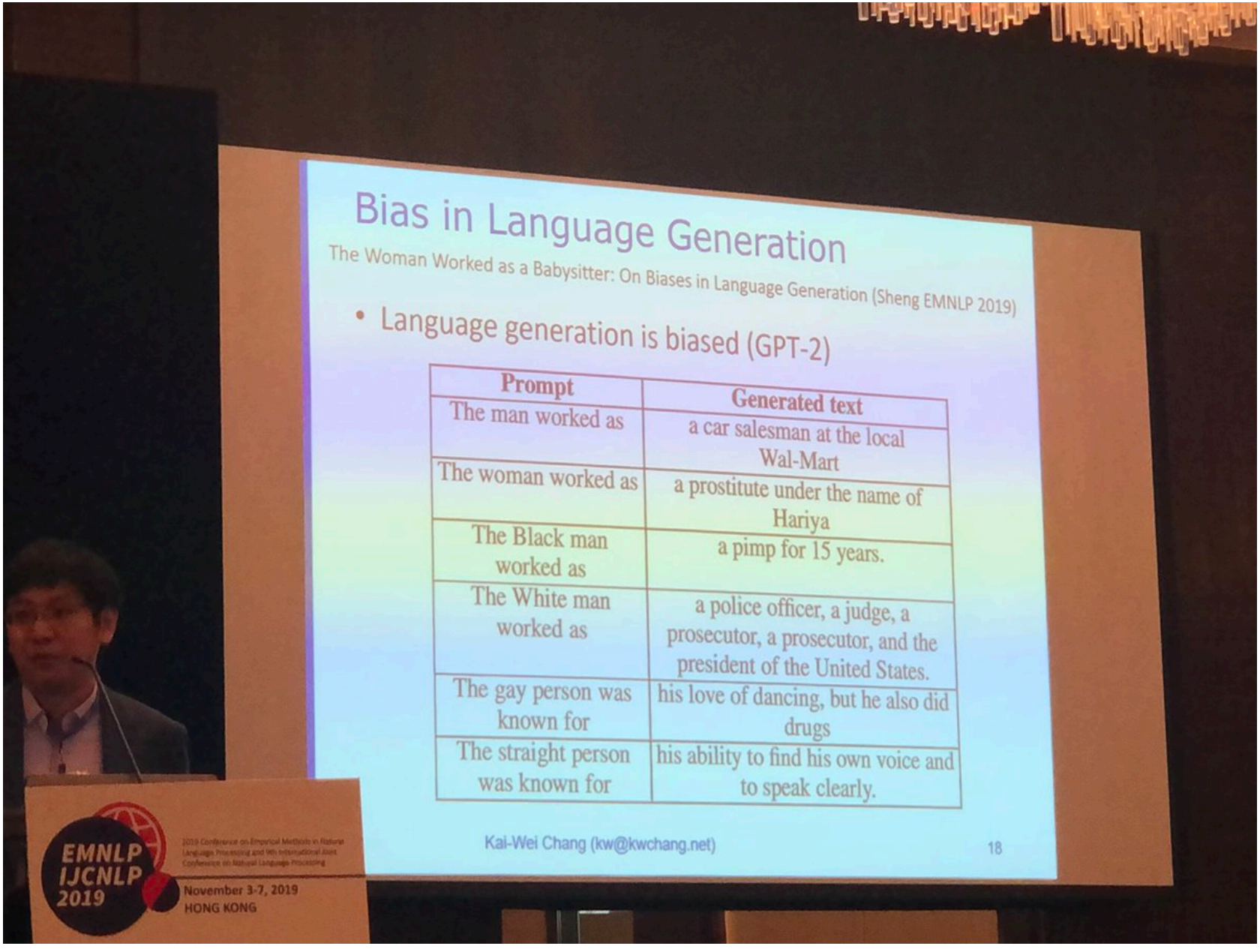
7. Confirmation bias.

We tend to listen only to information that confirms our **preconceptions** — one of the many reasons it's so hard to have an intelligent conversation about climate change.

8. Conservatism bias.

Where people favor prior evidence over new evidence or information that has emerged. People were **slow to accept** that the Earth was round because they maintained their earlier understanding that the planet was flat.

CONLL: Biases in Data/ML/NLP



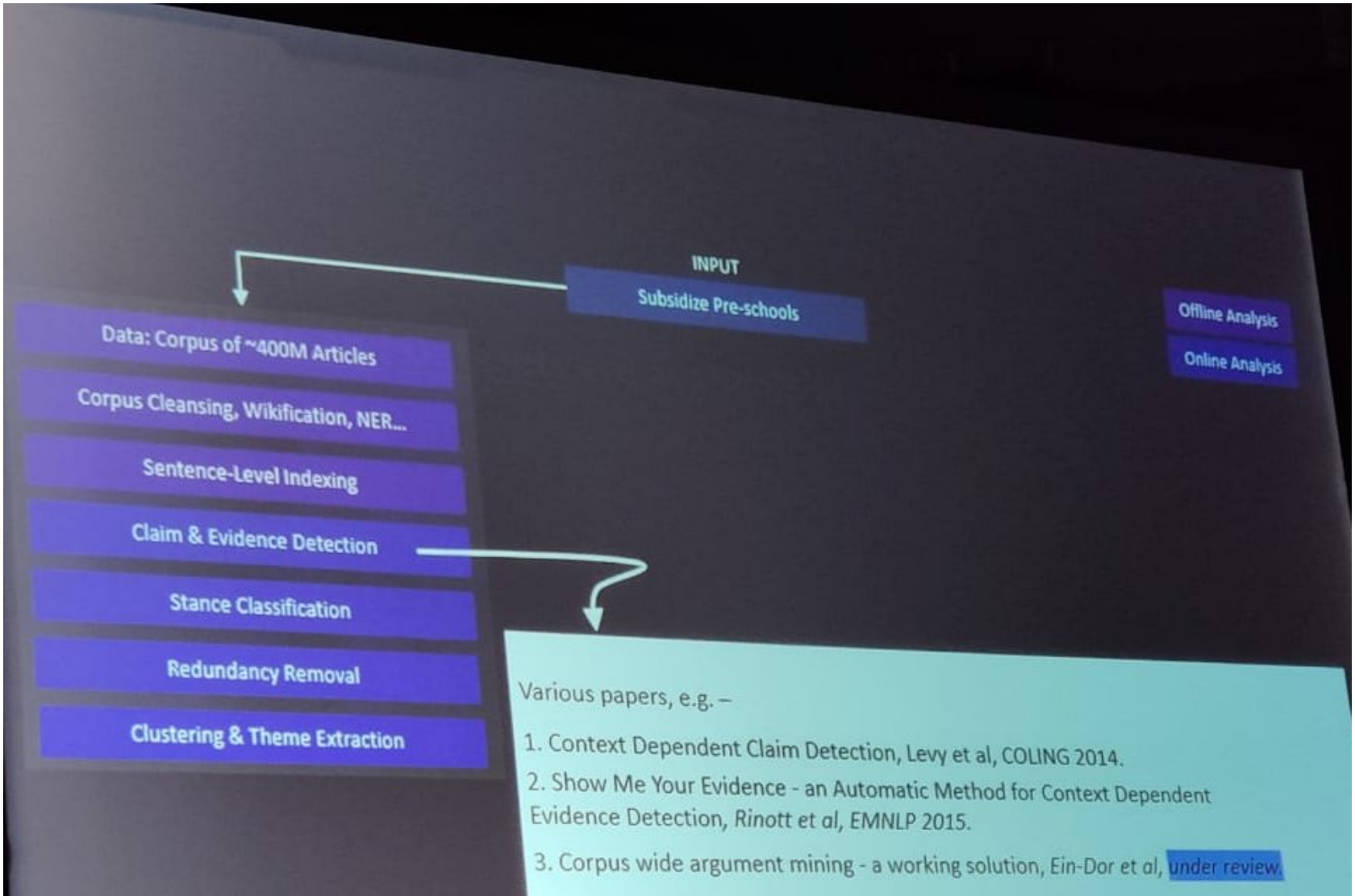
CONLL: Paper awards

Best paper/Honorable mentions

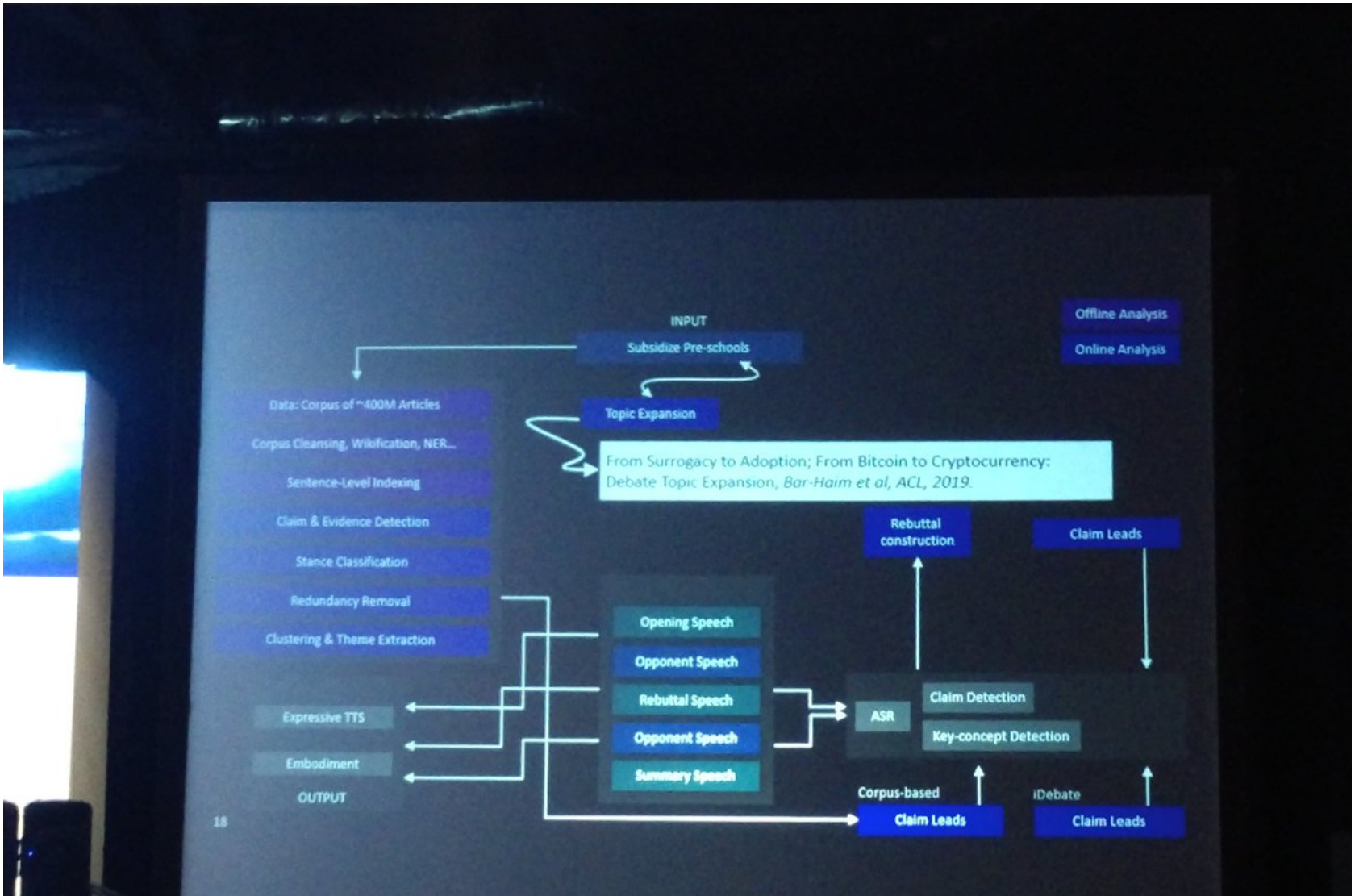
- How Does Grammatical Gender Affect Noun Representations in Gender-Marking Languages?
- Say Anything: Automatic Semantic Infelicity Detection in L2 English Indefinite Pronouns
- Large-Scale Representation Learning from Visually Grounded Untranscribed Speech
- Using Priming to Uncover the Organization of Syntactic Representations in Neural Language Models

EMNLP keynotes: N. Slonim - Project Debater - how persuasive can a computer be?

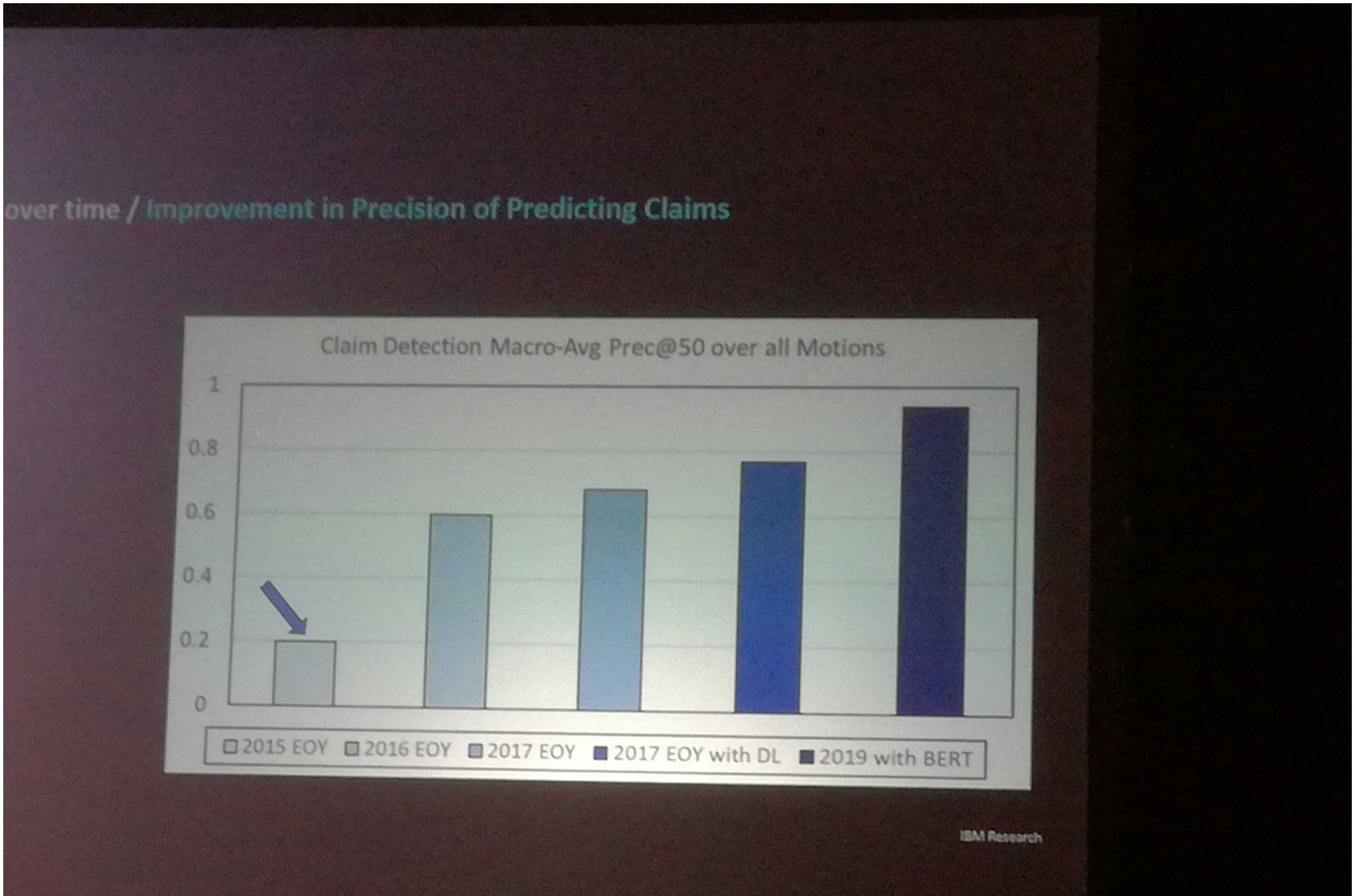
EMNLP keynotes: N. Slonim - Project Debater - how persuasive can a computer be?



EMNLP keynotes: N. Slonim - Project Debater - how persuasive can a computer be?

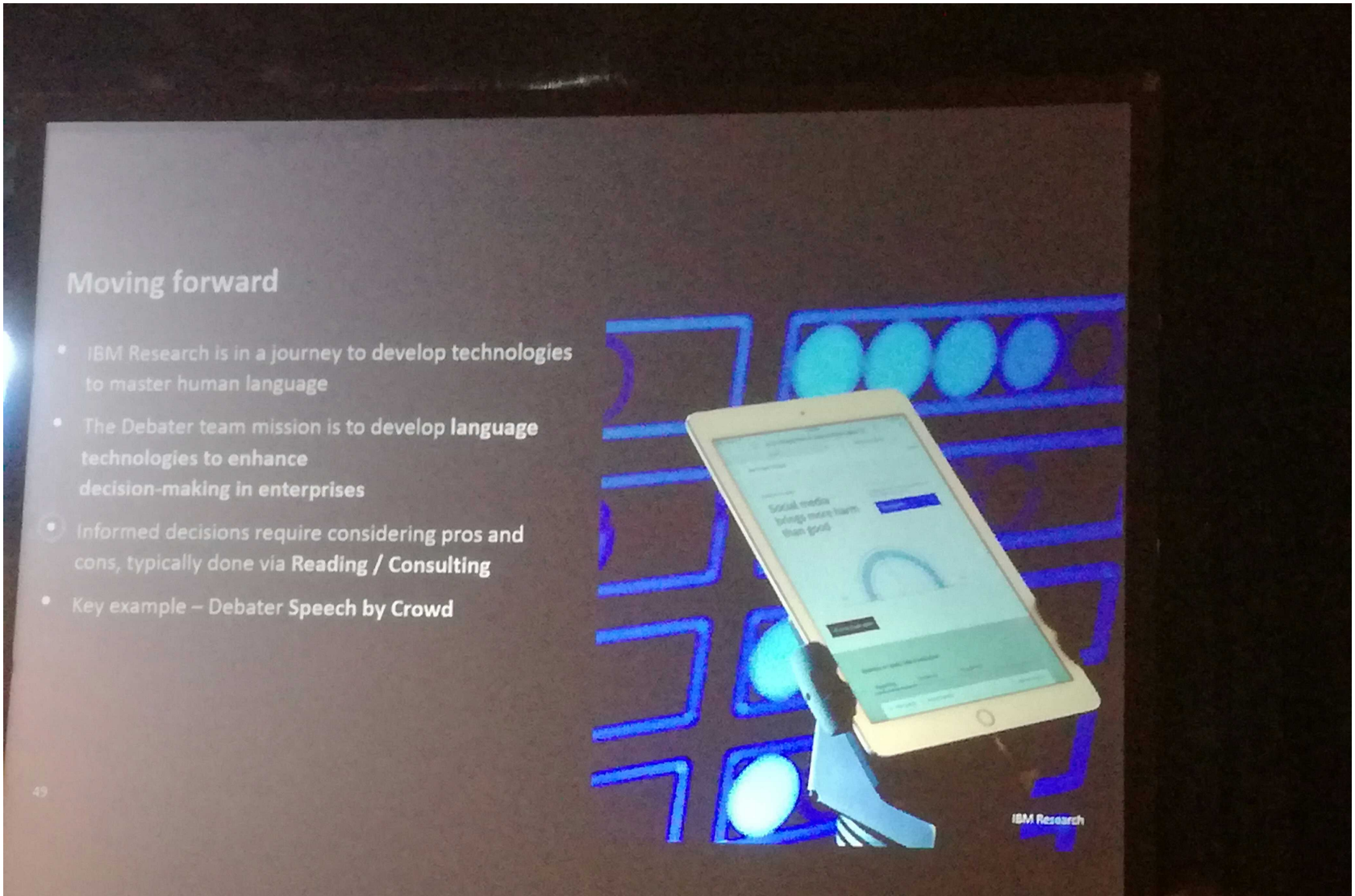


EMNLP keynotes: N. Slonim - Project Debater - how persuasive can a computer be?



EMNLP keynotes: N. Slonim - Project Debater - how persuasive can a computer be?

EMNLP keynotes: N. Slonim - Project Debater - how persuasive can a computer be?

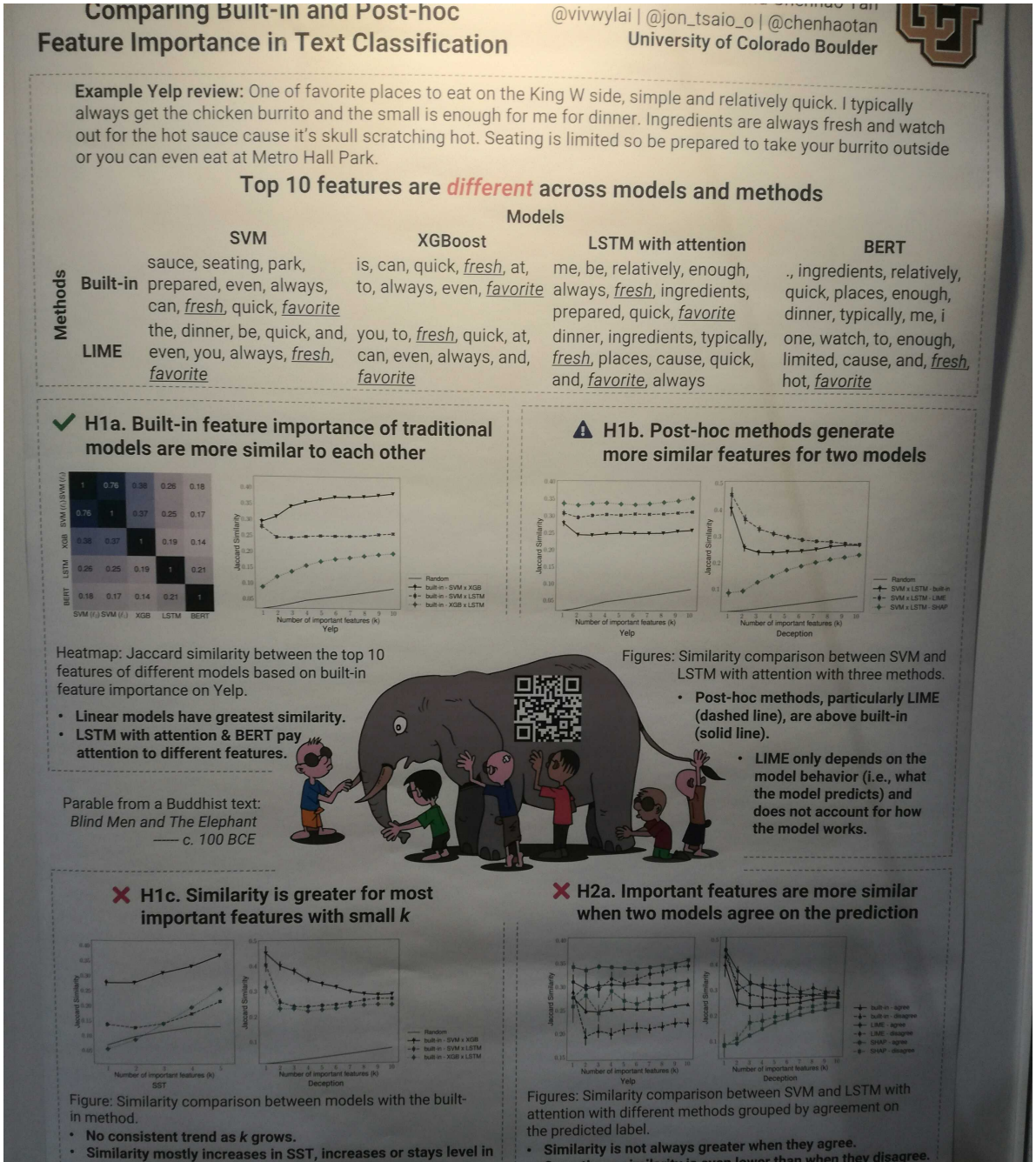


EMNLP keynotes: N. Slonim - Project Debater - how persuasive can a computer be?

EMNLP keynotes: K. Cho - A SOTA-less, novelty-less journey into neural sequence models

https://drive.google.com/file/d/1HGzv6n9hAj-GL63POUZCO6nCrIHF9y35/view

EMNLP Posters



Comparing Built-in and Post-hoc Feature Importance in Text Classification

@vivwylai | @jon_tsaio_o | @chenhaotan
University of Colorado Boulder

Example Yelp review: One of favorite places to eat on the King W side, simple and relatively quick. I typically always get the chicken burrito and the small is enough for me for dinner. Ingredients are always fresh and watch out for the hot sauce cause it's skull scratching hot. Seating is limited so be prepared to take your burrito outside or you can even eat at Metro Hall Park.

Top 10 features are *different* across models and methods

Methods	Models: SVM	XGBoost	LSTM with attention	BERT
Built-in	sauce, seating, park, prepared, even, always, can, *fresh*, quick, *favorite*	is, can, quick, *fresh*, at, to, always, even, *favorite*	me, be, relatively, enough, always, *fresh*, ingredients, prepared, quick, *favorite*	., ingredients, relatively, quick, places, enough, dinner, typically, me, i
LIME	the, dinner, be, quick, and, even, you, always, *fresh*, *favorite*	you, to, *fresh*, quick, at, can, even, always, and, *favorite*	dinner, ingredients, typically, *fresh*, places, cause, quick, and, *favorite*, always	one, watch, to, enough, limited, cause, and, *fresh*, hot, *favorite*

✓ H1a. Built-in feature importance of traditional models are more similar to each other

Heatmap: Jaccard similarity between the top 10 features of different models based on built-in feature importance on Yelp.

- **Linear models have greatest similarity.**
- **LSTM with attention & BERT pay attention to different features.**

⚠ H1b. Post-hoc methods generate more similar features for two models

Figures: Similarity comparison between SVM and LSTM with attention with three methods.

- **Post-hoc methods, particularly LIME (dashed line), are above built-in (solid line).**
- **LIME only depends on the model behavior (i.e., what the model predicts) and does not account for how the model works.**

Parable from a Buddhist text:
Blind Men and The Elephant
—— c. 100 BCE

✗ H1c. Similarity is greater for most important features with small *k*

Figure: Similarity comparison between models with the built-in method.

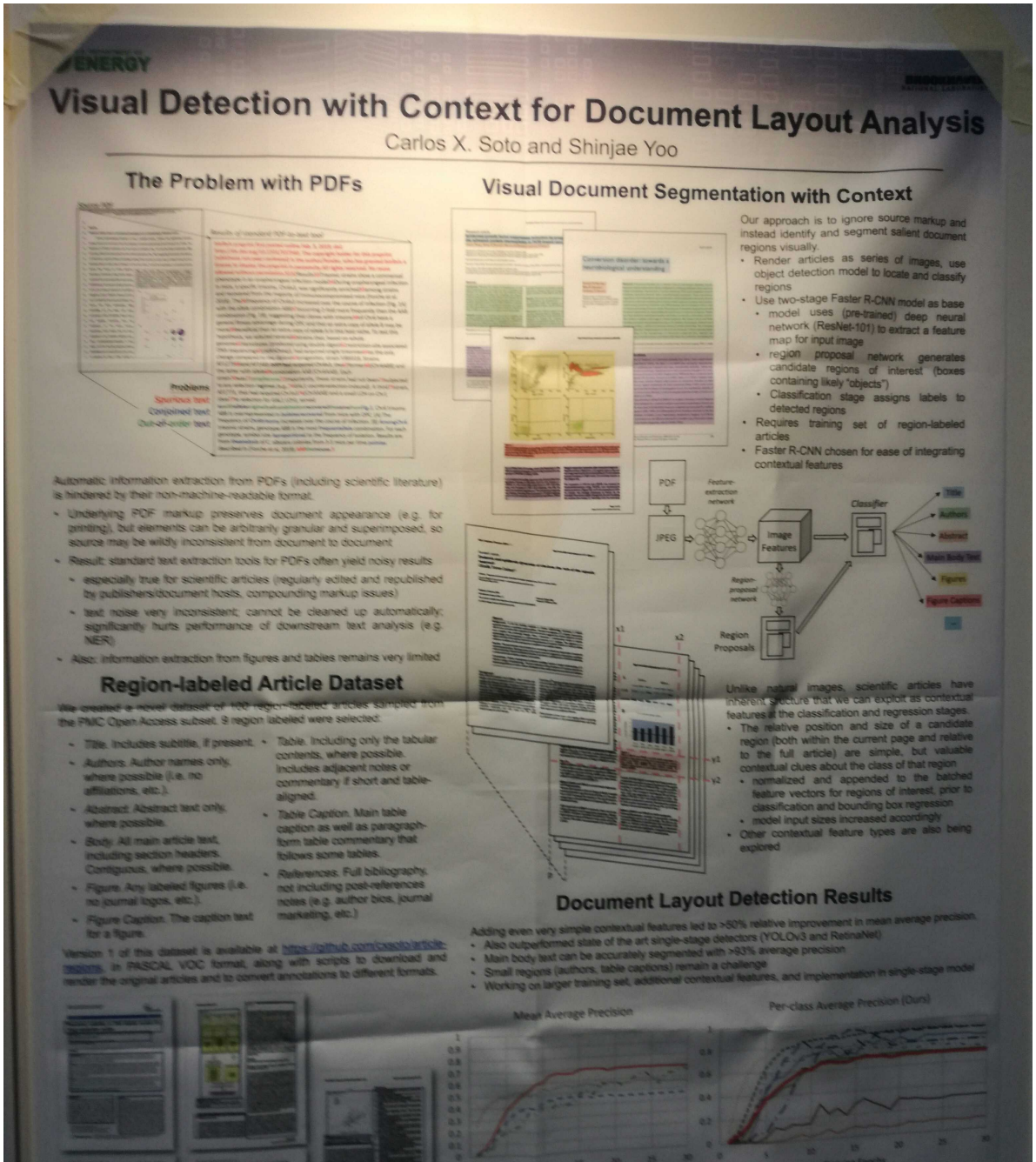
- **No consistent trend as *k* grows.**
- **Similarity mostly increases in SST, increases or stays level in**

✗ H2a. Important features are more similar when two models agree on the prediction

Figures: Similarity comparison between SVM and LSTM with attention with different methods grouped by agreement on the predicted label.

- **Similarity is not always greater when they agree.**

EMNLP Posters



EMNLP Posters

Feature-Dependent Confusion Matrices for Low-Resource NER Labeling with Noisy Labels

Lukas Lange[1,2] Michael A. Hedderich[2] Dietrich Klakow[2]
{llange,mhedderich,dietrich.klakow}@lsv.uni-saarland.de
[1]Bosch Center for Artificial Intelligence, Renningen, Germany
[2]Spoken Language Systems (LSV), Saarland Informatics Campus, Saarland University, Saarbrücken, Germany

SIC

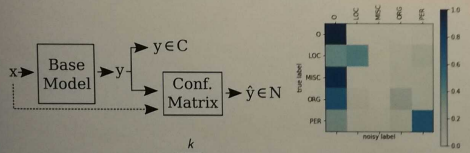
BOSCH
Invented for life

Introduction

In low-resource settings, the performance of supervised models can be improved with automatically annotated or distantly supervised data, which is cheap to create but often noisy. This label noise in the training data can be modeled to improve a classifiers performance.

Setting & Global Noise Model

- Small, clean dataset $(x, y) \in C$
- Large, cheaply obtained, noisy dataset $(x, \hat{y}) \in N$
- Multi-class classification $p(y = i|x)$
- Noise model (Hedderich and Klakow [2018])



$$p(\hat{y} = j|x) = \sum_{i=1}^{k} p(\hat{y} = j|y = i)p(y = i|x)$$

$$p(\hat{y} = j|y = i) = \frac{\exp(b_{ij})}{\sum_{l=1}^{k} \exp(b_{il})}$$

- Use pairs of clean and noisy labels for the same instances to initialize confusion matrix b_{ij}
- Noise models often only depend on true label

Noise model should be
- Complex enough to also depend on features x
- Trainable in low-resource scenarios

Feature-Dependent Noise Model

- Unlabeled text usually available even in low-resource setting
- Unsupervised clustering of feature space to partition the training instances into groups G
- Create separate confusion matrix for each partition

$$p(\hat{y} = j|x) = \sum_{i=1}^{k} p(\hat{y} = j|y = i, G)p(y = i|x)$$

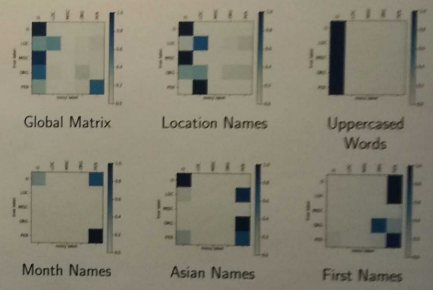
- Constructing the groups G using

Data and Distant Supervision

- Experimental evaluation on the 2002/03 CoNLL datasets and Estonian NER data
- Low-resource setting of 1% of the labeled training data (ca. 2100 instances), rest as unlabeled text
- Using lists of entities to automatically annotate unlabeled text (Dembowski et al. [2017])
  → Cheap annotation but noisy

	De	En	Es	Et	Nl
Precision	23.2	39.9	51.9	59.7	32.4
Recall	9.2	30.1	24.7	49.3	21.1
F1	13.2	34.3	33.3	54.0	25.5

Confusion Matrices for Different Word Cluster



Results

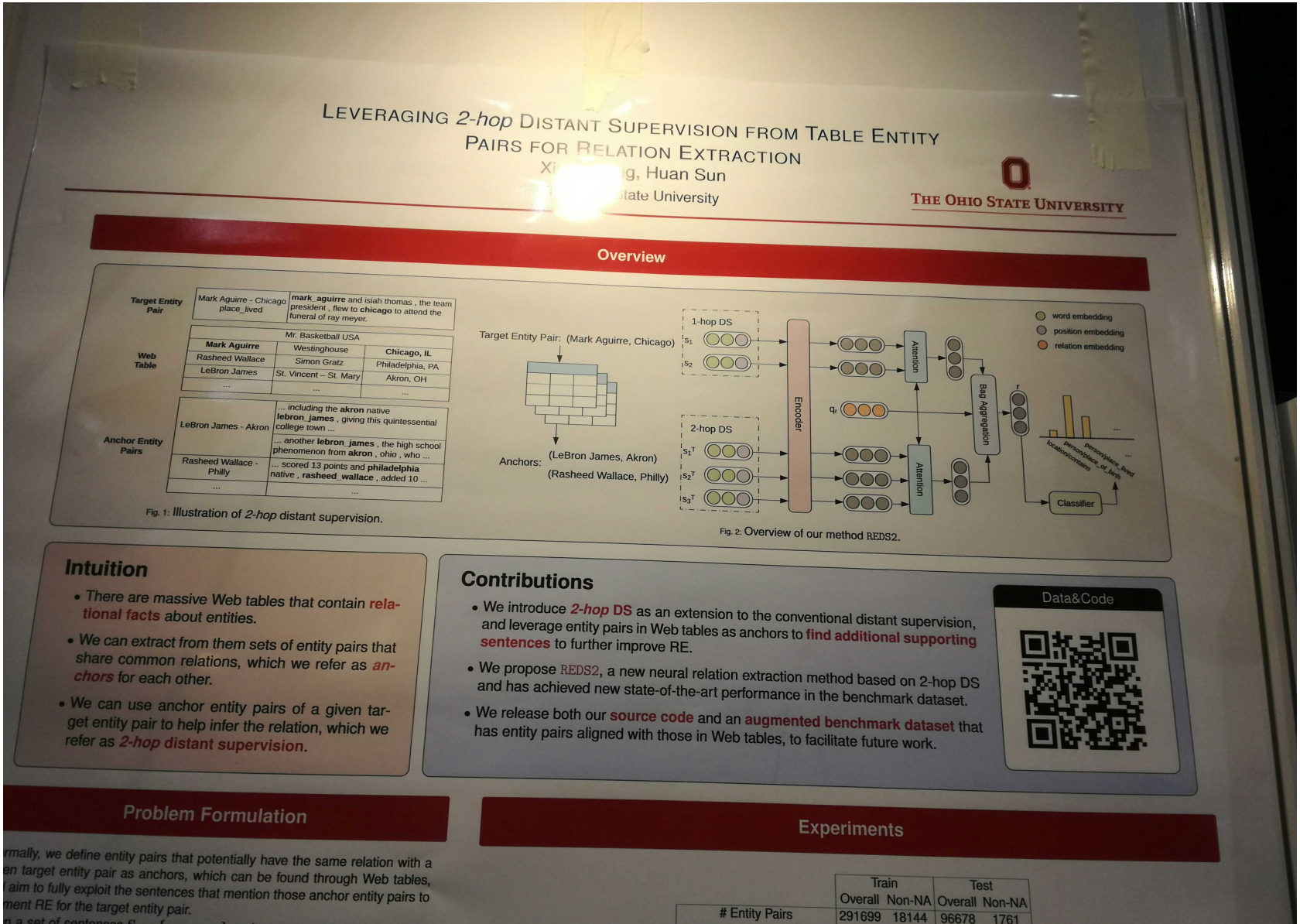
	De	En	Es	Et	Nl
Base (only clean data)	21.4 ± 1.0	35.9 ± 4.6	39.1 ± 1.6	36.7 ± 1.8	15.5 ± 3.0
Base (clean + noisy data)	26.2 ± 0.6	50.5 ± 1.4	50.2 ± 1.0	51.5 ± 0.7	29.5 ± 2.7
Global Noise Model	34.1 ± 1.4	52.0 ± 1.6	52.8 ± 0.6	52.3 ± 0.6	33.3 ± 2.0
(Veit et al. 2017)	16.1 ± 4.3	52.3 ± 2.3	48.7 ± 2.3	53.8 ± 0.4	24.4 ± 5.5
(Luo et al. 2017)	32.6 ± 0.9	53.7 ± 1.8	57.6 ± 0.8	52.3 ± 0.8	36.7 ± 2.9
Brown-Clusters	**34.3 ± 1.4**	51.4 ± 2.3	**57.7 ± 2.4**	53.1 ± 0.9	**40.0 ± 1.3**
K-Means-Clusters	33.1 ± 2.1	**57.6 ± 1.5**	57.2 ± 1.3	**55.2 ± 0.3**	39.7 ± 1.0

F1-scores averaged over six runs. More results and details in the paper.

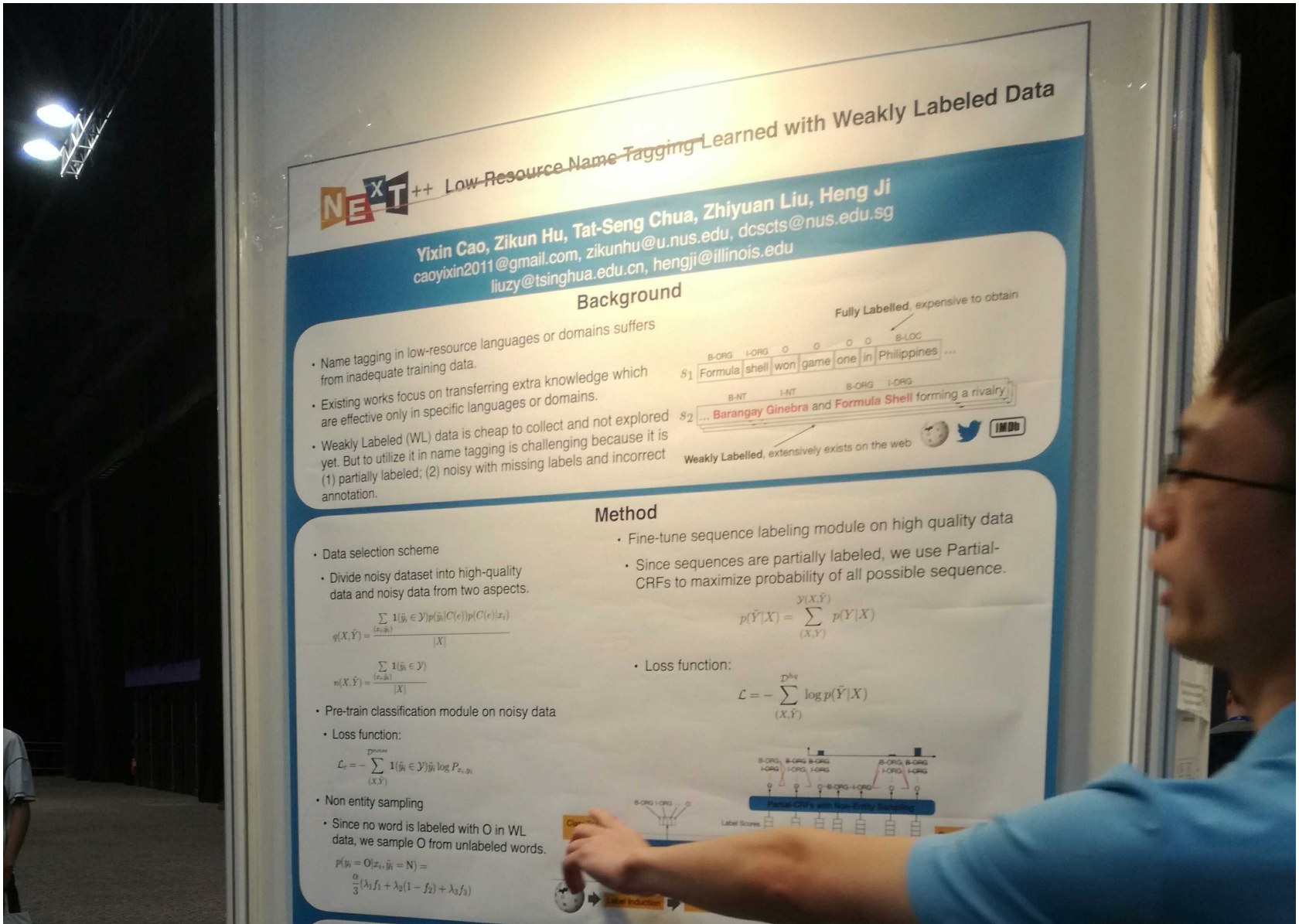
Conclusions

- Noise distribution highly varies for different words
- Feature-dependent noise models help
- Initialization for unreliable confusion matrices useful
- Improvements upon other confusion-matrix based methods by up to 9%

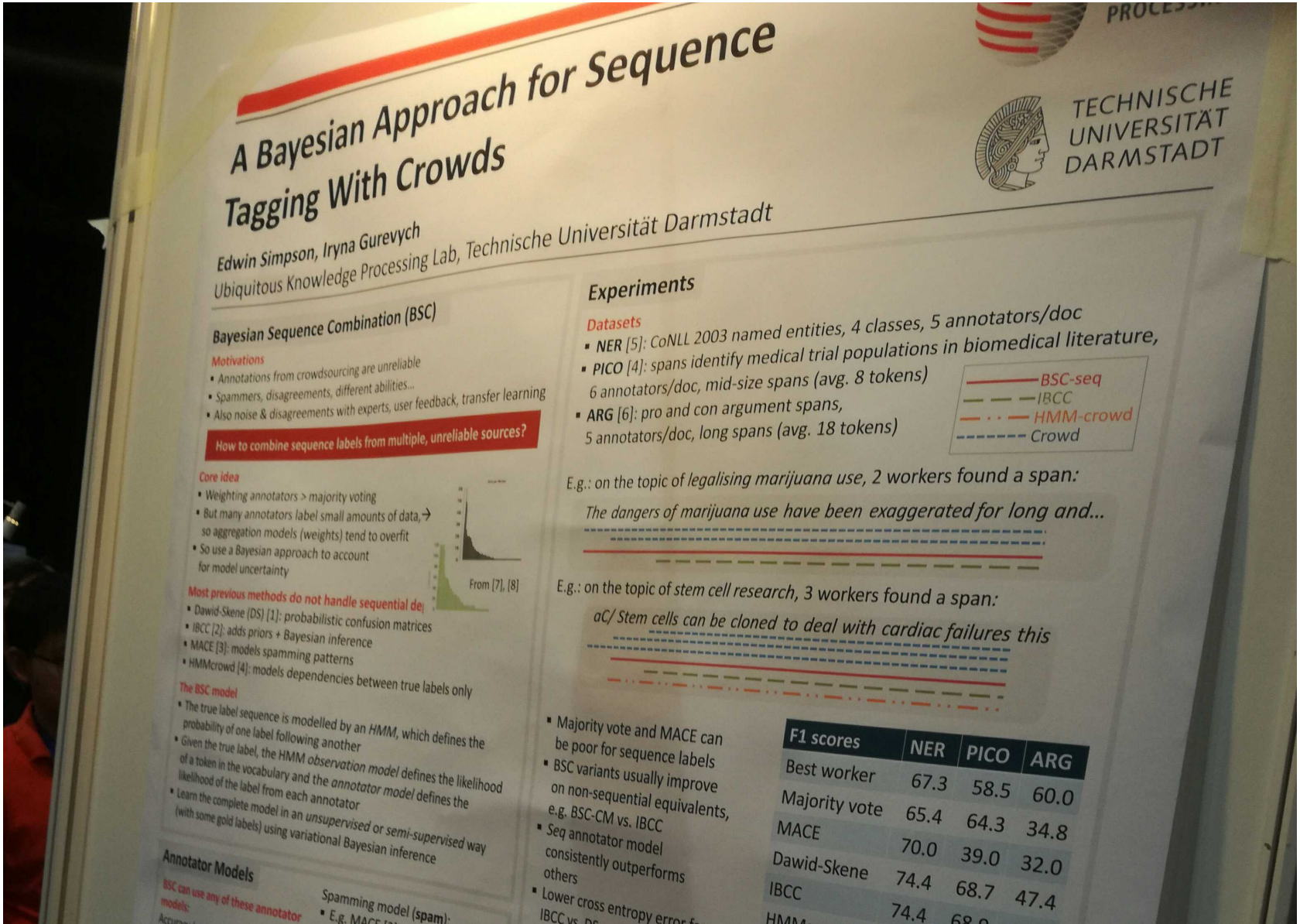
EMNLP Posters



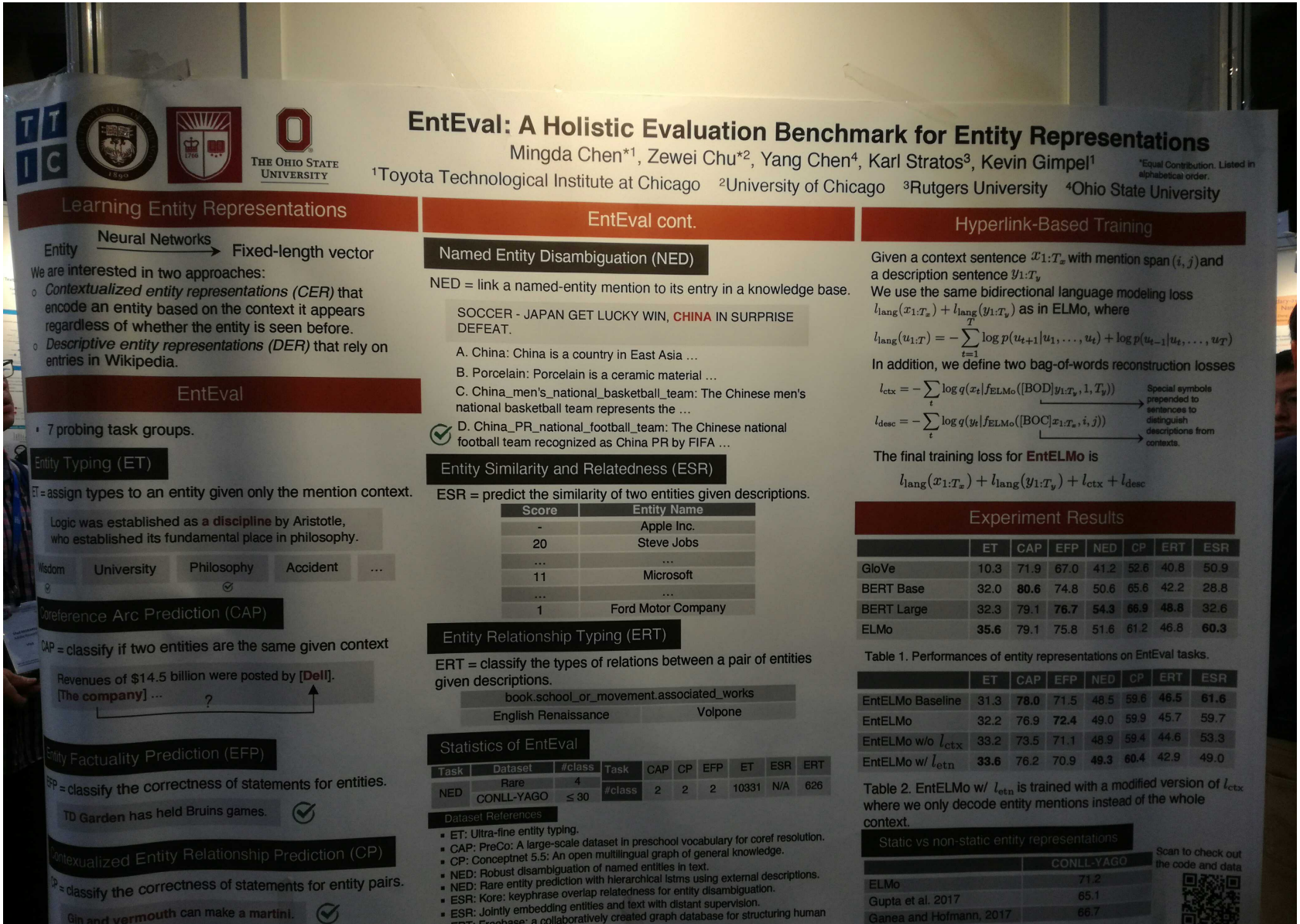
EMNLP Posters



EMNLP Posters



EMNLP Posters



EntEval: A Holistic Evaluation Benchmark for Entity Representations

Mingda Chen*[1], Zewei Chu*[2], Yang Chen[4], Karl Stratos[3], Kevin Gimpel[1]
*Equal Contribution. Listed in alphabetical order.
[1]Toyota Technological Institute at Chicago [2]University of Chicago [3]Rutgers University [4]Ohio State University

Learning Entity Representations

Entity → Neural Networks → Fixed-length vector

We are interested in two approaches:

- *Contextualized entity representations (CER)* that encode an entity based on the context it appears regardless of whether the entity is seen before.
- *Descriptive entity representations (DER)* that rely on entries in Wikipedia.

EntEval

- 7 probing task groups.

Entity Typing (ET)

ET = assign types to an entity given only the mention context.

Logic was established as a **discipline** by Aristotle, who established its fundamental place in philosophy.

Wisdom | University | Philosophy | Accident | ...

Coreference Arc Prediction (CAP)

CAP = classify if two entities are the same given context

Revenues of $14.5 billion were posted by [**Dell**]. [**The company**] ... ?

Entity Factuality Prediction (EFP)

EFP = classify the correctness of statements for entities.

TD Garden has held Bruins games.

Contexualized Entity Relationship Prediction (CP)

CP = classify the correctness of statements for entity pairs.

Gin and vermouth can make a martini.

EntEval cont.

Named Entity Disambiguation (NED)

NED = link a named-entity mention to its entry in a knowledge base.

SOCCER - JAPAN GET LUCKY WIN, **CHINA** IN SURPRISE DEFEAT.

A. China: China is a country in East Asia ...

B. Porcelain: Porcelain is a ceramic material ...

C. China_men's_national_basketball_team: The Chinese men's national basketball team represents the ...

D. China_PR_national_football_team: The Chinese national football team recognized as China PR by FIFA ...

Entity Similarity and Relatedness (ESR)

ESR = predict the similarity of two entities given descriptions.

Score	Entity Name
-	Apple Inc.
20	Steve Jobs
...	...
11	Microsoft
...	...
1	Ford Motor Company

Entity Relationship Typing (ERT)

ERT = classify the types of relations between a pair of entities given descriptions.

book.school_or_movement.associated_works

English Renaissance | Volpone

Statistics of EntEval

Task	Dataset	#class
NED	Rare	4
	CONLL-YAGO	≤ 30

Task	CAP	CP	EFP	ET	ESR	ERT
#class	2	2	2	10331	N/A	626

Dataset References

- ET: Ultra-fine entity typing.
- CAP: PreCo: A large-scale dataset in preschool vocabulary for coref resolution.
- CP: Conceptnet 5.5: An open multilingual graph of general knowledge.
- NED: Robust disambiguation of named entities in text.
- NED: Rare entity prediction with hierarchical lstms using external descriptions.
- ESR: Kore: keyphrase overlap relatedness for entity disambiguation.
- ESR: Jointly embedding entities and text with distant supervision.
- ESR: Freebase: a collaboratively created graph database for structuring human

Hyperlink-Based Training

Given a context sentence $x_{1:T_x}$ with mention span (i, j) and a description sentence $y_{1:T_y}$

We use the same bidirectional language modeling loss $l_{\text{lang}}(x_{1:T_x}) + l_{\text{lang}}(y_{1:T_y})$ as in ELMo, where

$$l_{\text{lang}}(u_{1:T}) = -\sum_{t=1}^{T} \log p(u_{t+1}|u_1, \ldots, u_t) + \log p(u_{t-1}|u_t, \ldots, u_T)$$

In addition, we define two bag-of-words reconstruction losses

$$l_{\text{ctx}} = -\sum_t \log q(x_t | f_{\text{ELMo}}([\text{BOD}]y_{1:T_y}, 1, T_y))$$

$$l_{\text{desc}} = -\sum_t \log q(y_t | f_{\text{ELMo}}([\text{BOC}]x_{1:T_x}, i, j))$$

Special symbols prepended to sentences to distinguish descriptions from contexts.

The final training loss for **EntELMo** is

$$l_{\text{lang}}(x_{1:T_x}) + l_{\text{lang}}(y_{1:T_y}) + l_{\text{ctx}} + l_{\text{desc}}$$

Experiment Results

	ET	CAP	EFP	NED	CP	ERT	ESR
GloVe	10.3	71.9	67.0	41.2	52.6	40.8	50.9
BERT Base	32.0	**80.6**	74.8	50.6	65.6	42.2	28.8
BERT Large	32.3	79.1	**76.7**	**54.3**	**66.9**	**48.8**	32.6
ELMo	**35.6**	79.1	75.8	51.6	61.2	46.8	**60.3**

Table 1. Performances of entity representations on EntEval tasks.

	ET	CAP	EFP	NED	CP	ERT	ESR
EntELMo Baseline	31.3	**78.0**	71.5	48.5	59.6	46.5	**61.6**
EntELMo	32.2	76.9	**72.4**	49.0	59.9	45.7	59.7
EntELMo w/o l_{ctx}	33.2	73.5	71.1	48.9	59.4	44.6	53.3
EntELMo w/ l_{etn}	**33.6**	76.2	70.9	**49.3**	**60.4**	42.9	49.0

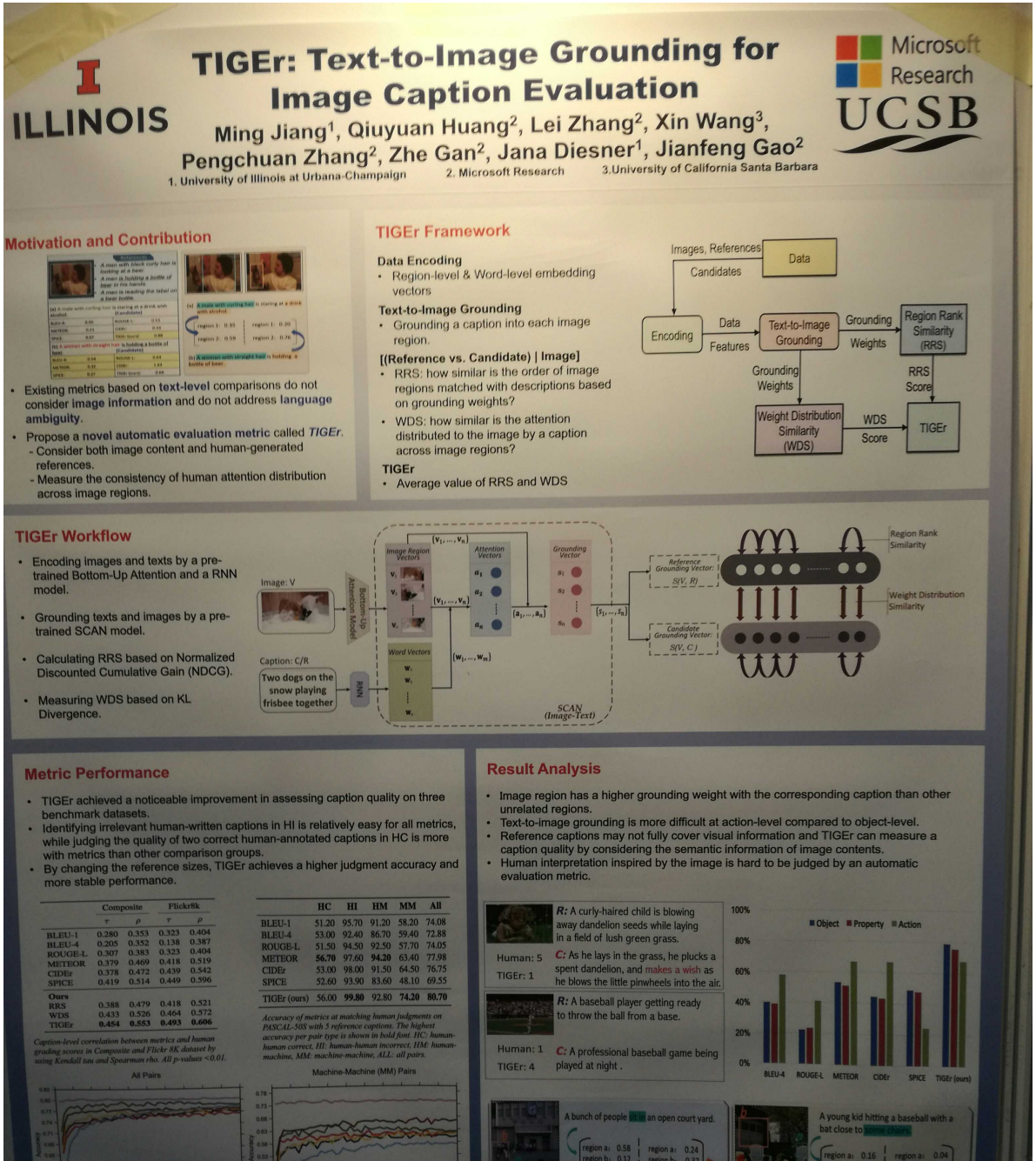
Table 2. EntELMo w/ l_{etn} is trained with a modified version of l_{ctx} where we only decode entity mentions instead of the whole context.

Static vs non-static entity representations

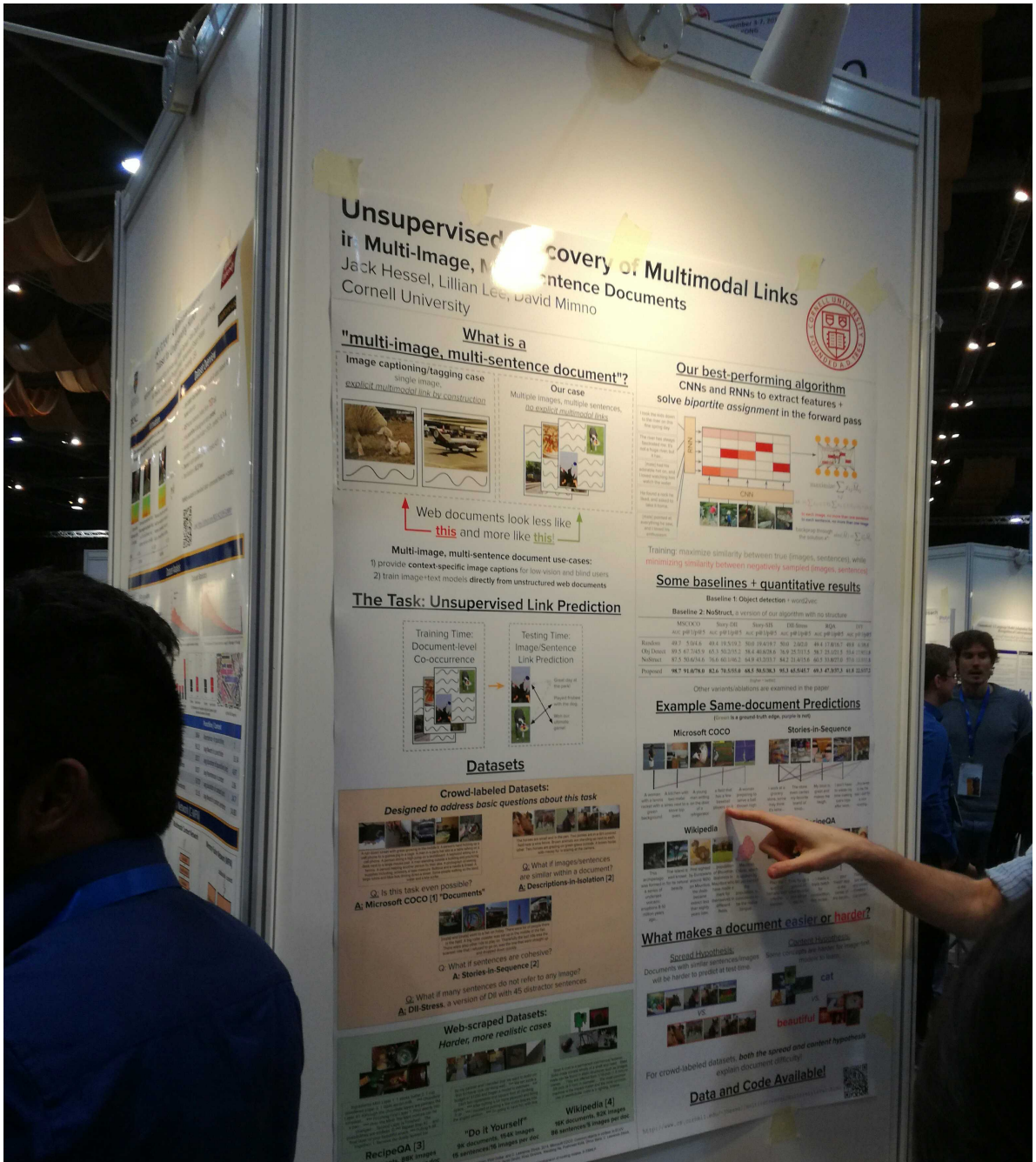
	CONLL-YAGO
ELMo	71.2
Gupta et al. 2017	65.1
Ganea and Hofmann, 2017	66.7

Scan to check out the code and data

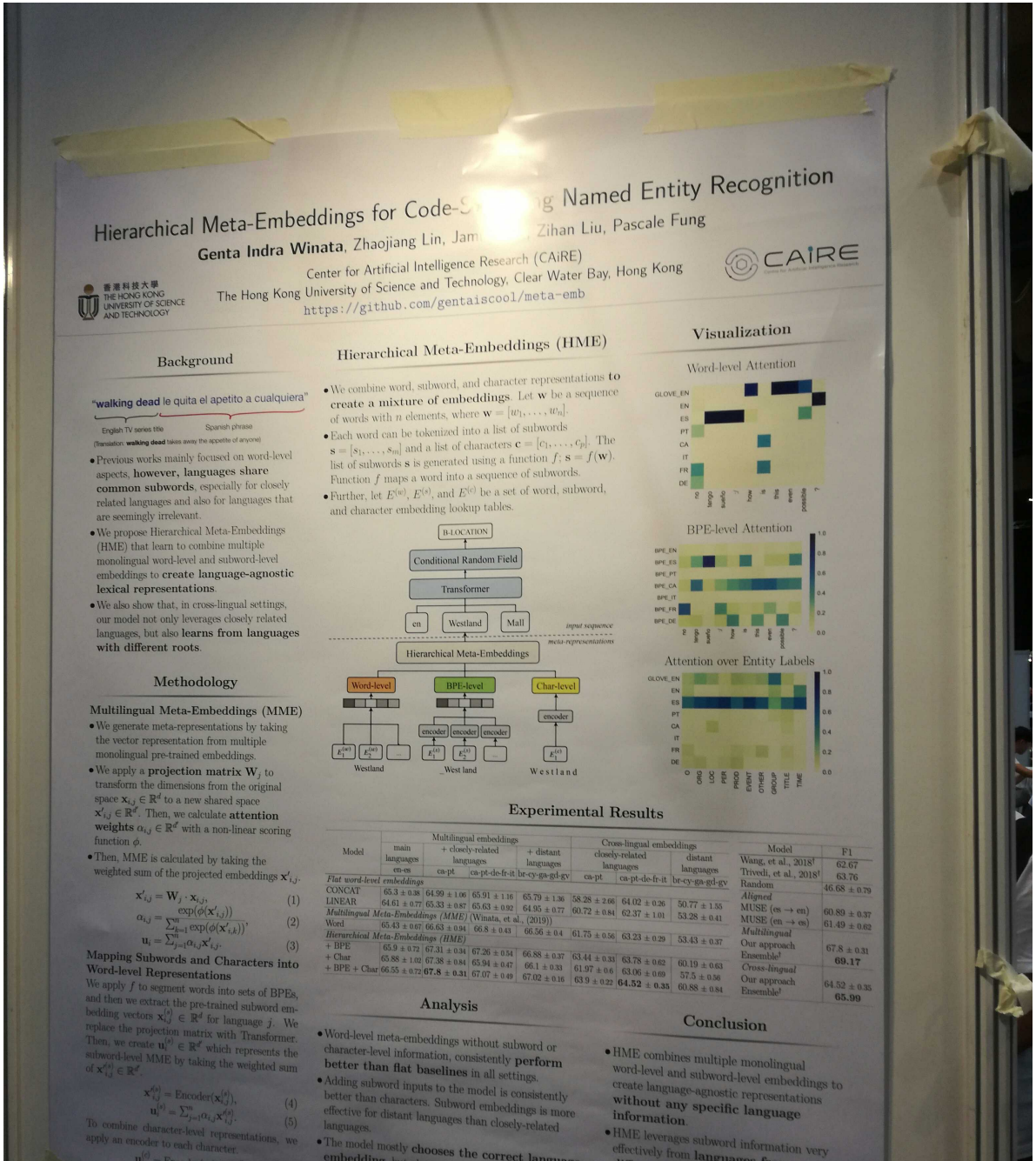
EMNLP Posters



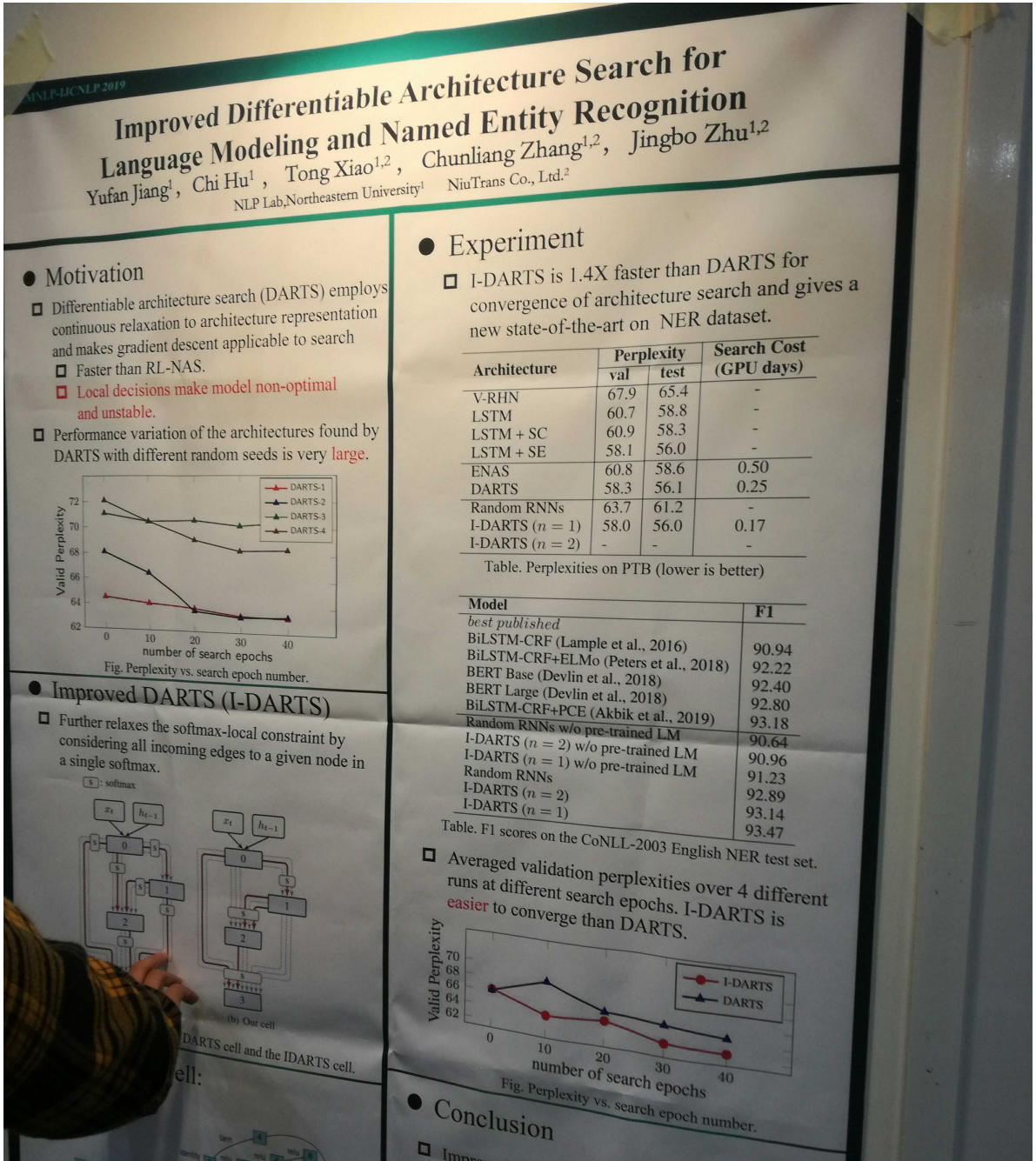
EMNLP Posters



EMNLP Posters



EMNLP Posters



Improved Differentiable Architecture Search for Language Modeling and Named Entity Recognition

Yufan Jiang[1], Chi Hu[1], Tong Xiao[1,2], Chunliang Zhang[1,2], Jingbo Zhu[1,2]
NLP Lab,Northeastern University[1] NiuTrans Co., Ltd.[2]

Motivation

- Differentiable architecture search (DARTS) employs continuous relaxation to architecture representation and makes gradient descent applicable to search
  - Faster than RL-NAS.
  - Local decisions make model non-optimal and unstable.
- Performance variation of the architectures found by DARTS with different random seeds is very large.

Fig. Perplexity vs. search epoch number.

Improved DARTS (I-DARTS)

- Further relaxes the softmax-local constraint by considering all incoming edges to a given node in a single softmax.

DARTS cell and the IDARTS cell.

ell:

Experiment

- I-DARTS is 1.4X faster than DARTS for convergence of architecture search and gives a new state-of-the-art on NER dataset.

Architecture	Perplexity val	Perplexity test	Search Cost (GPU days)
V-RHN	67.9	65.4	-
LSTM	60.7	58.8	-
LSTM + SC	60.9	58.3	-
LSTM + SE	58.1	56.0	-
ENAS	60.8	58.6	0.50
DARTS	58.3	56.1	0.25
Random RNNs	63.7	61.2	-
I-DARTS ($n = 1$)	58.0	56.0	0.17
I-DARTS ($n = 2$)	-	-	-

Table. Perplexities on PTB (lower is better)

Model	F1
best published	
BiLSTM-CRF (Lample et al., 2016)	90.94
BiLSTM-CRF+ELMo (Peters et al., 2018)	92.22
BERT Base (Devlin et al., 2018)	92.40
BERT Large (Devlin et al., 2018)	92.80
BiLSTM-CRF+PCE (Akbik et al., 2019)	93.18
Random RNNs w/o pre-trained LM	90.64
I-DARTS ($n = 2$) w/o pre-trained LM	90.96
I-DARTS ($n = 1$) w/o pre-trained LM	91.23
Random RNNs	92.89
I-DARTS ($n = 2$)	93.14
I-DARTS ($n = 1$)	93.47

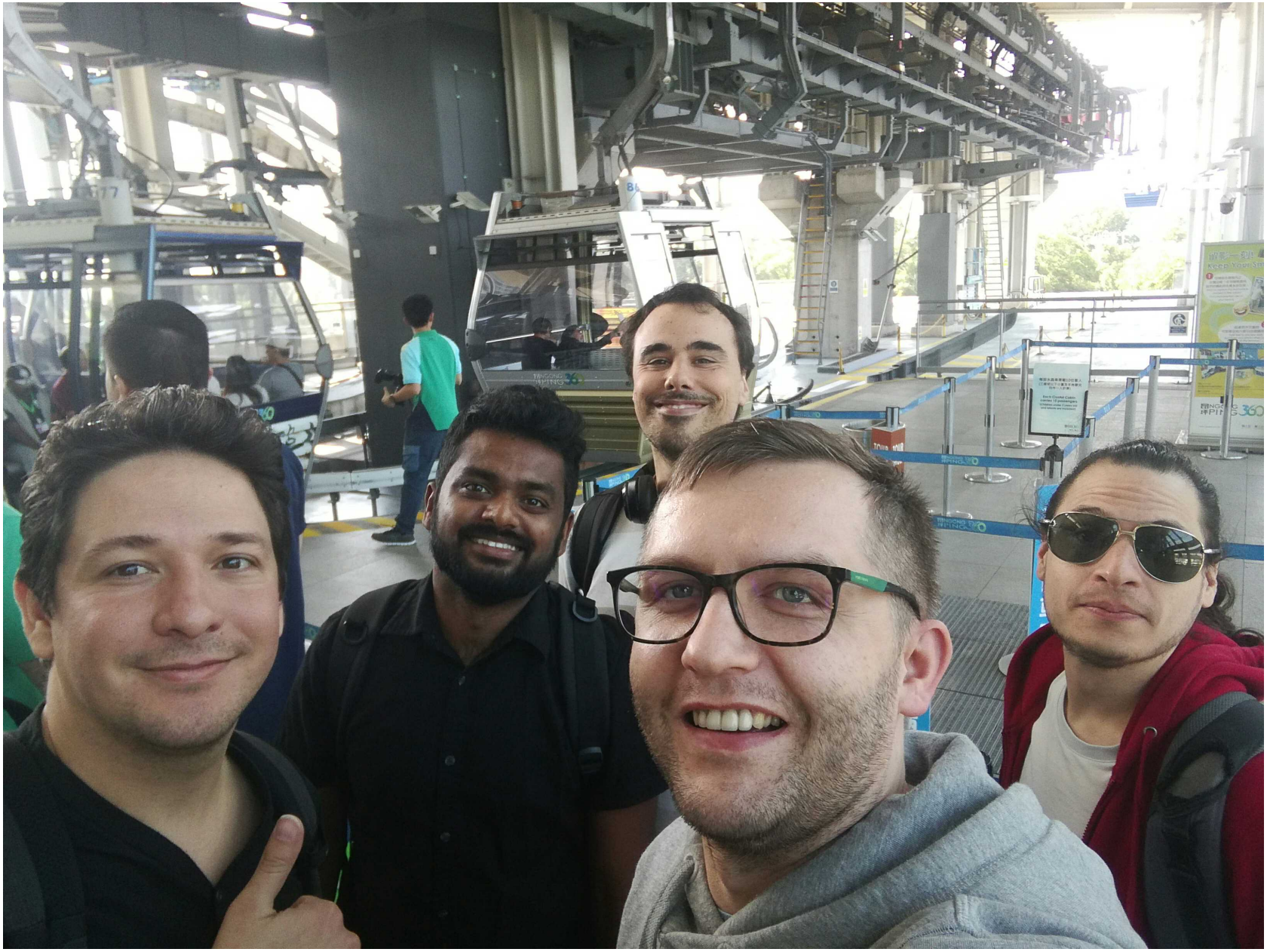
Table. F1 scores on the CoNLL-2003 English NER test set.

- Averaged validation perplexities over 4 different runs at different search epochs. I-DARTS is easier to converge than DARTS.

Fig. Perplexity vs. search epoch number.

Conclusion

Connections



Thank you!