

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

TL;DR

Tomasz Stanislawek

Aplica 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)

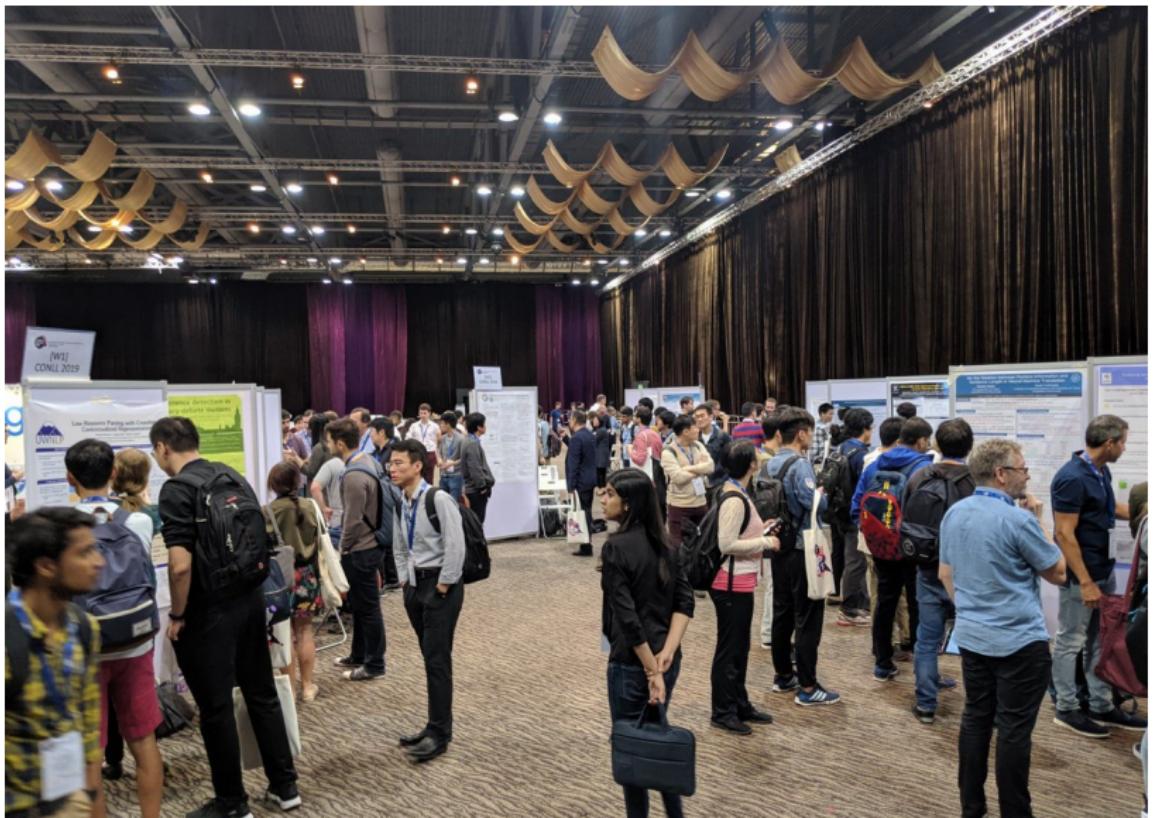
CONLL/Workshops



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)

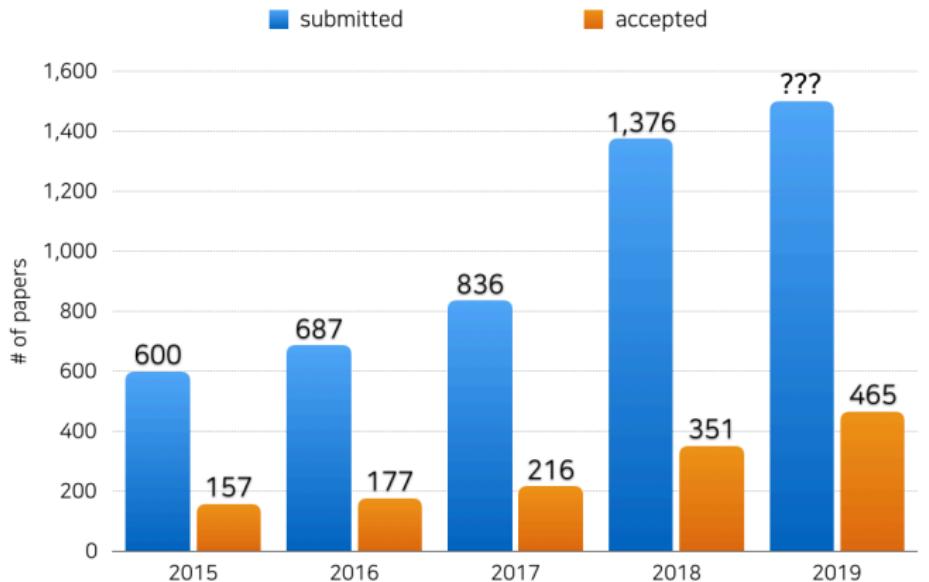
CONLL/Workshops





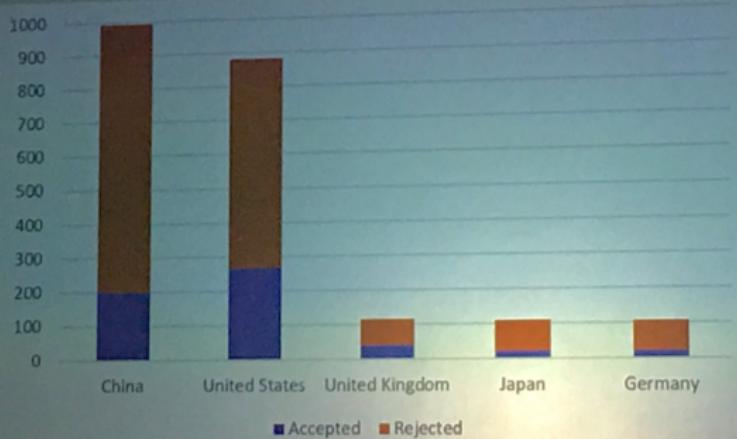
ACCEPTANCE STATISTICS

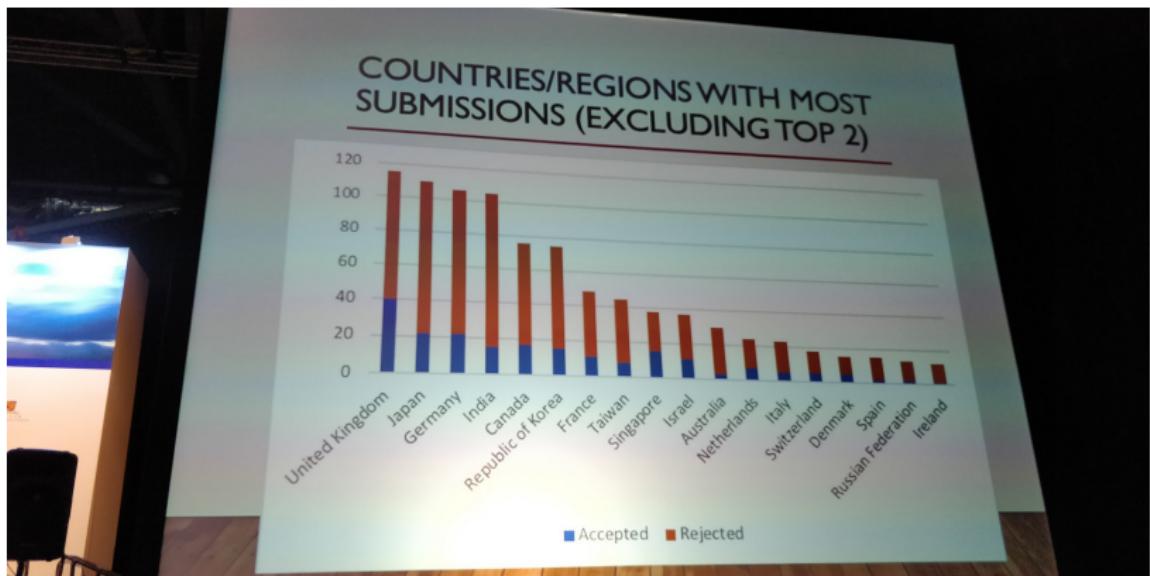
	Long Papers	Short Papers	Total
Reviewed	1813	1063	2876
Accepted as Oral	164	48	212
Accepted as Poster	301	170	471
Total Accepted	465 (25.6%)	218 (20.5%)	683 (23.7%)





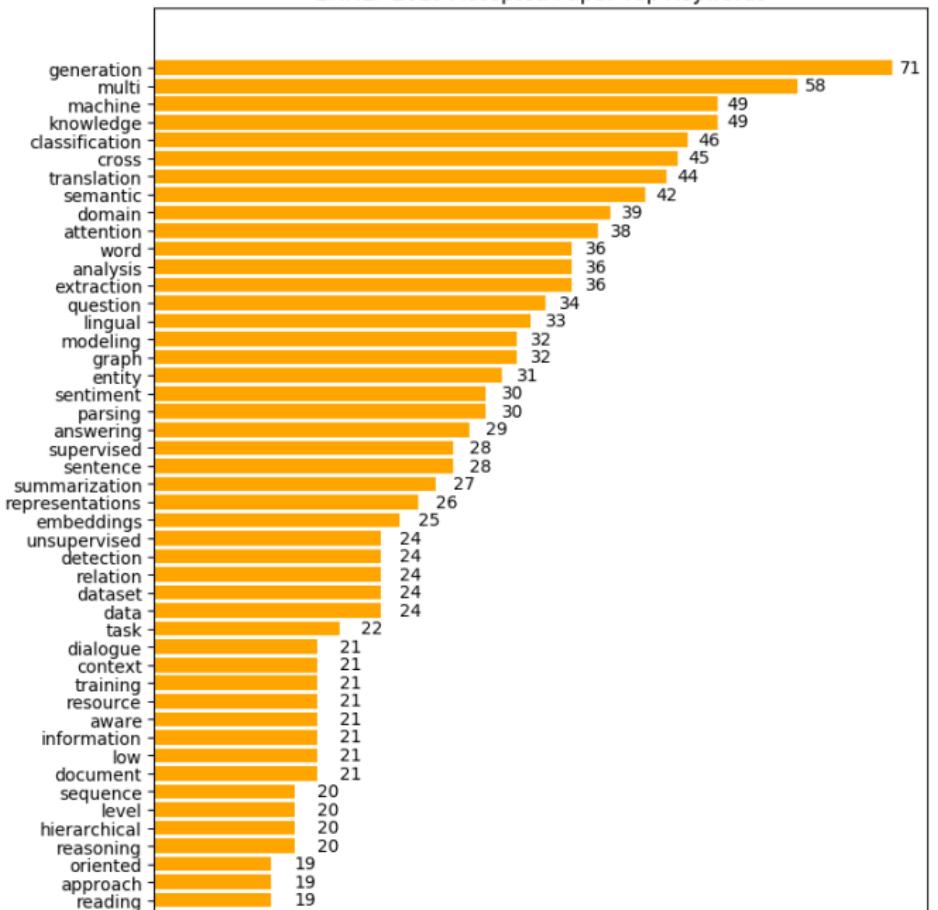
COUNTRIES/REGIONS WITH MOST SUBMISSIONS (TOP 5)







EMNLP 2019 Accepted Paper Top Keywords





Country or Region	All submissions			Long submissions			Short submissions		
	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	83	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



Two problems

BERT and friends are **awesome** as a universal pre-training base
Nevertheless, we still have work to do:

1. We've built powerful, **neural matching machines**, rather than devices that can **think**
2. We've build devices for one-step classification, QA etc., rather than devices that can **reason** through a series of steps

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



Effective use of vast amounts of visual data



Improving Human Computer Interaction



Challenging multi-modal AI research problem



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**?

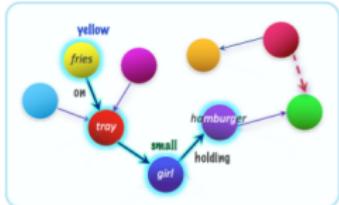
CONLL: C. Manning - Multi-step reasoning for answering complex questions



Pattern: What/Which <type> {do you think} <is> <object>, <attr> or <decoy>
Program: Select: <object> → 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

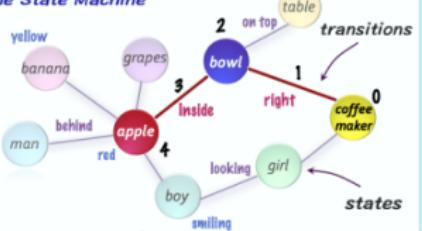
- Consistency
- Validity & Plausibility
- Distribution
- Grounding

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



What is the red fruit inside the bowl to the right of the coffee maker?

The State Machine



instructions

alphabet (concepts)



properties

disentangled representation

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



What is the **tall** object to the **left** of the **bed** made of?



Cabinet: **wood** (0.95), **tall** (0.92), **shiny** (0.86)
Bed: **white** (0.84), **comfortable** (0.91)
Lamp: **yellow** (0.92), **on** (0.74), **thin** (0.82)

→ **Wood**

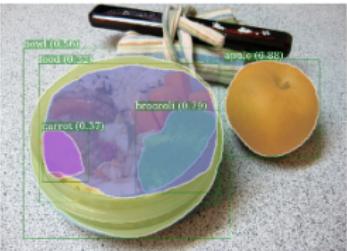


What is the **green** food **inside** of the **bowl**?



Apple: **yellow** (0.58), **round** (0.95), **healthy** (0.91)
Broccoli: **green** (0.94), **leafy** (0.93), **fresh** (0.92)
Bowl: **plastic** (0.72), **transparent** (0.84)

→ **Broccoli**



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



Let's build neural networks that think!
By seeking tasks involving
understanding and
multi-step compositional reasoning

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



Thank you!

Relevant papers:

Drew Hudson and Christopher Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. CVPR.

Drew Hudson and Christopher Manning. 2019. Learning by Abstraction: The Neural State Machine. arXiv:1907.03950. NeurIPS.

Zhilin Yang*, Peng Qi*, Saizheng Zhang*, Yoshua Bengio, William W. Cohen, Ruslan Salakutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. EMNLP.

Peng Qi, Vera Lin*, Leo Mehr*, Zijian Wang*, and Christopher D. Manning. 2019. Answering Complex Open-domain Questions Through Iterative Query Generation. EMNLP.

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.koccon|piotr.milkowski}@pwr.edu.pl)

Monika Zaśko-Zielńska
(monika.zasko-zielinska@uwr.edu.pl)

Wrocław University
of Science and Technology

Uniwersytet
Wrocławski

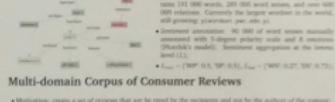
Main Objectives

- 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 two distance measures;
- 5. Evaluate generalization ability of models tested in cross-domain setting.

Key Contribution

- Detailed description of the procedure of building PolEmo 2.0: manually annotated corpus of reviews or news from 4 domains (domains, vehicles, health, books, products) in 2 levels of granularity (document, sentence);
- Detailed analysis of manual annotation with regard to frequently occurring errors;
- Performance results obtained on a new dataset (PolEmo 2.0), compared with state-of-the-art approaches (SST-2 corpus, also using sentiment lexicons derived from PolEmo 1.0).
- Performance results (F1 score) evaluated on texts within a given domain, (M2) evaluated on texts from different domains, (M3) evaluated on texts from all domains, (M4) evaluated on texts from a single domain and (M5) evaluated on texts from all domains (M6) evaluated on texts from all domains with respect to domain independence.
- Comparison of classic (Logistic Regression) and deep learning models, mainly in the context of the ability to generate the sentiment recognition problem and providing domain-independent sentiment measures;
- Making PolEmo 2.0 available under an open license: <https://github.com/jkoccon/PolEmo2.0>

Sentiment Dictionary



Multi-domain Corpus of Consumer Reviews

- Motivation: create a set of reviews that can be used by the researchers not only at the level of the entire review, but also at the sentence level;
- Goal 1: the motivation 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: books (tripadvisor.com), medicine (medscape.com), vehicles (polbase.pl), products (eurosyn.pl), 20% news (polbase.pl).
- Annotation type: (NP) merely positive, (NPV) positive with some less relevant negative aspects, (NN) neutral, (NNV) negative with some less relevant positive aspects, (NNV) entirely negative, (AMR) these are both positive and negative aspects in the text that are balanced in terms of relevance.

Positive Specific Agreement

	D	SP	NP	NN	AMR	AMR	NP	A			
H	91.91	36.29	41.39	39.38	91.63	90.73	11.61	99.42	78.58	91.32	89.39
P	94.85	33.33	10.67	47.62	85.95	73.68	78.76	94.85	85.95	94.85	85.95
F	94.85	20.89	6.67	34.07	92.34	54.19	77.62	95.85	89.07	94.85	89.07
S	87.50	20.89	6.67	34.00	92.34	54.19	77.63	95.50	86.00	93.22	88.32

Evaluation Sets

Type	Domain	Total	Dev	Test	Holdout	Others	Demands	SP	AMR	N	NP
SST	Medicine	2818	327	327	1272		Medicine	29.37	38.37	24.11	36.95
SST	Products	387	49	49	494		Product	31.16	27.48	30.84	34.65
SST	School	403	50	51	304		School	27.44	36.47	4.46	27.46
M2T	All	6523	825	825	2486		Medicine	29.33	32.26	42.24	35.55
SST	Books	18126	2264	2264	22667		Medicine	23.08	36.26	35.95	31.98
SST	Products	5142	743	742	7427		Product	24.43	18.98	18.36	44.62
SST	School	2024	250	251	2021		School	26.77	26.77	26.77	26.77
M2T	All	26744	3745	3747	37466		All	26.67	11.98	24.24	34.95

Table 5. The number of items according to the domain.

• 1D – Single Domain – evaluation was carried using elements from the same domain;

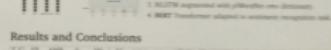
• 2D – Domain (the – review) was created using elements from 2 domains, one per the remaining domain. It allows to evaluate the ability of the method to capture the domain-independent sentiment features;

• M2T – Mixed Domains – 1D – each item was rated independently. This version allows to measure the ability of the classifier to generalize the task of sentiment analysis on all available domains.

Multi-level Text Classification

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

Approaches



Results and Conclusions

Task	C	SP	AMR	N	NP	Demands	SP	AMR	N	NP	
SP	1.33	51.55	86.80	86.77	97.20	94.28	73.36	45.05	64.44	85.95	94.46
AMR	1.33	87.39	91.90	89.85	94.44	94.44	72.82	53.72	82.72	89.78	95.78
N	4.33	39.19	91.92	91.92	91.92	91.92	74.94	4.92	94.47	94.47	78.28
NP	2.78	38.18	91.92	91.92	91.92	91.92	74.94	4.92	94.47	94.47	78.28
Demands	3.03	31.13	54.67	47.79	74.73	74.73	92.70	76.32	94.46	94.46	82.17
SP	1.33	74.87	91.92	91.92	91.92	91.92	74.94	29.23	76.32	72.08	70.32
AMR	1.33	74.87	91.92	91.92	91.92	91.92	74.94	29.23	76.32	72.08	70.32
N	7.62	41.77	49.85	49.85	73.54	73.54	74.94	27.24	81.82	73.54	76.72
NP	2.78	20.89	46.81	46.81	73.54	73.54	74.94	22.22	74.73	73.54	76.72
Demands	4.33	41.77	49.85	49.85	73.54	73.54	74.94	36.36	80.77	73.54	76.72
SP	4.33	87.39	91.92	91.92	91.92	91.92	74.94	45.05	64.44	85.95	94.46
AMR	4.33	41.77	49.85	49.85	73.54	73.54	74.94	42.29	74.73	73.54	76.72
N	1.33	51.55	86.80	86.77	97.20	94.28	73.36	45.05	64.44	85.95	94.46
NP	1.33	87.39	91.92	91.92	91.92	91.92	74.94	4.92	94.47	94.47	78.28
Demands	3.03	31.13	54.67	47.79	74.73	74.73	92.70	14.87	94.46	94.46	82.17

•

CONLL: J. Kocon - 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



Training data are
collected and
annotated

Human Biases in Data		
Reporting bias	Stereotypical bias	Group attribution error
Selection bias	Historical unfairness	Halo effect
Overgeneralization	Implicit associations	
Out-group homogeneity bias	Implicit stereotypes	
	Prejudice	
Human Biases in Collection and Annotation		
Sampling error	Bias blind spot	Neglect of probability
Non-sampling error	Confirmation bias	Anecdotal fallacy
Insensitivity to sample size	Subjective validation	Illusion of validity
Correspondence bias	Experimenter's bias	
In-group bias	Choice-supportive bias	

CONLL: Biases in Data/ML/NLP



Biases in Data

Selection Bias: Selection does not reflect a random sample

- Men are over-represented in web-based news articles
(Jia, Lansdall-Welfare, and Cristianini 2015)
- Men are over-represented in twitter conversations
(Garcia, Weber, and Garimella 2014)
- Gender bias in Wikipedia and Britannica
(Reagle & Rhuee 2011)

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



Bias in Language Generation

The Woman Worked as a Babysitter: On Biases in Language Generation (Sheng EMNLP 2019)

- Language generation is biased (GPT-2)

Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

Kai-Wei Chang (kw@kwchang.net)

18

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?



From Checkers to Debate and beyond...

From Checkers to Chess & Go in ~70 years -

All in the 'comfort zone' of AI -

- Easy to know who won
- Can define a scoring function to quantify any state
- Computer can win via tactics humans do not comprehend

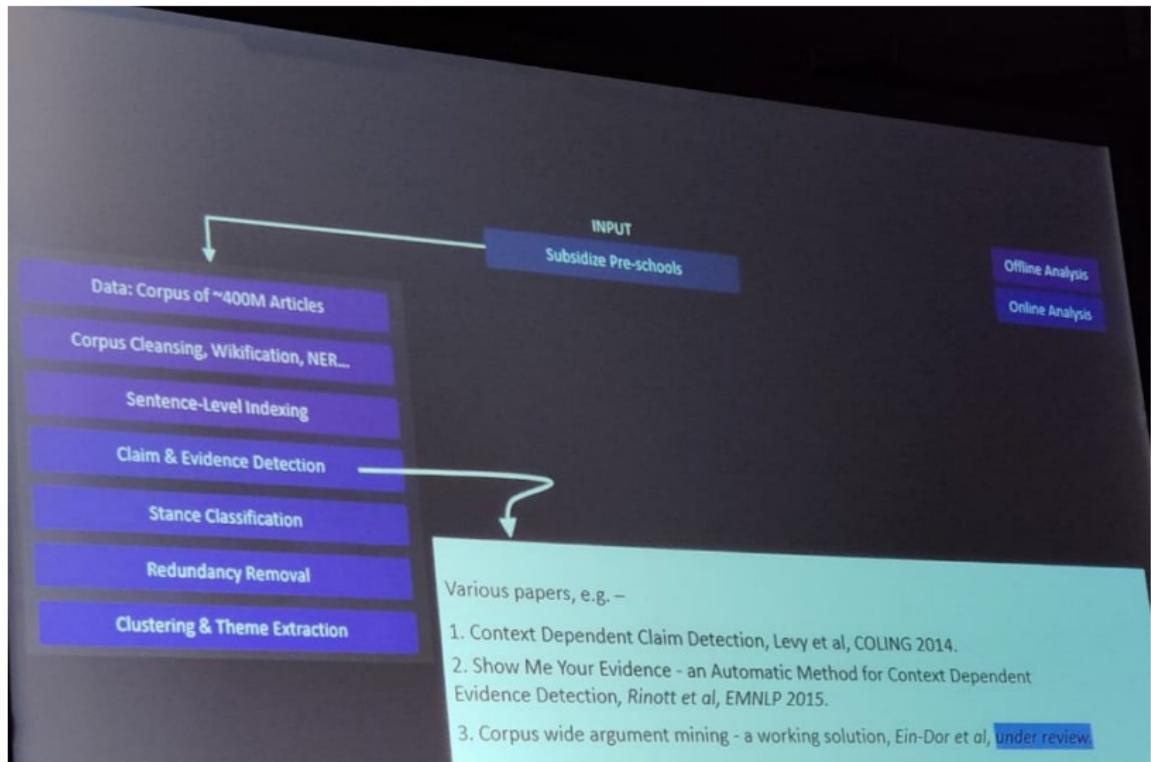


→ A new territory for AI
Grand Challenges?

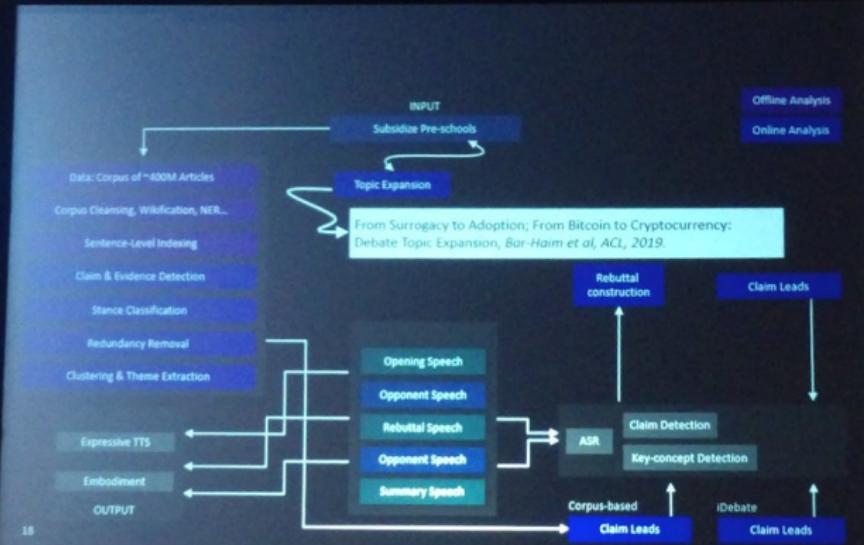


IBM Research

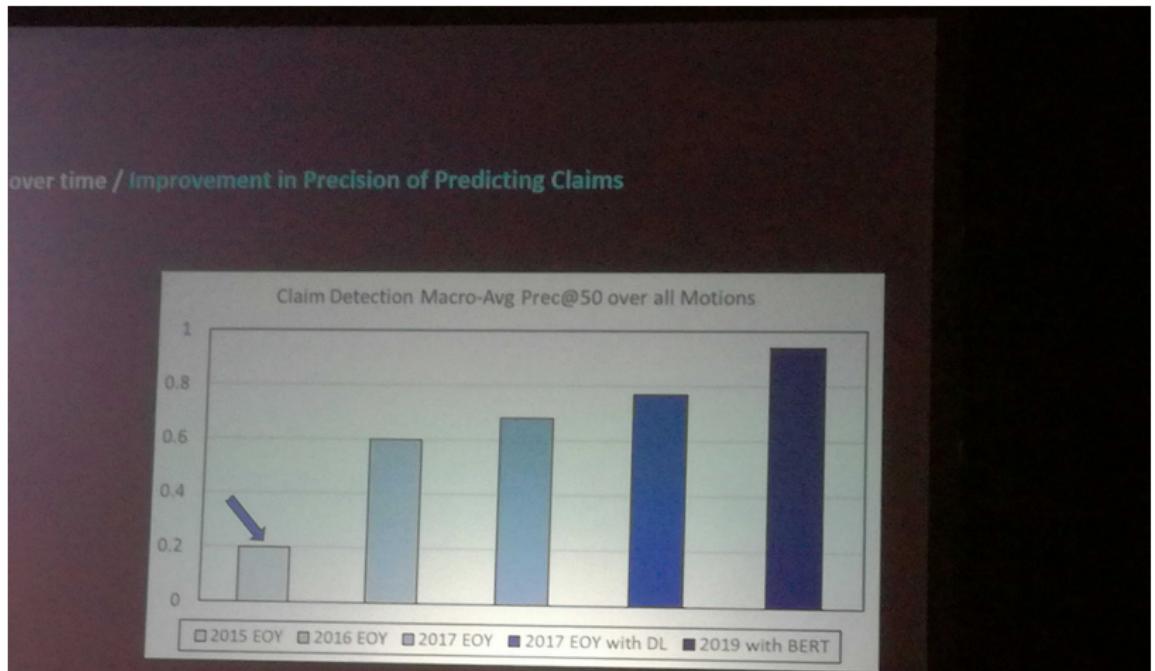
EMNLP keynotes: N. Slonim - Project Debater - how persuasive can a computer be?



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?



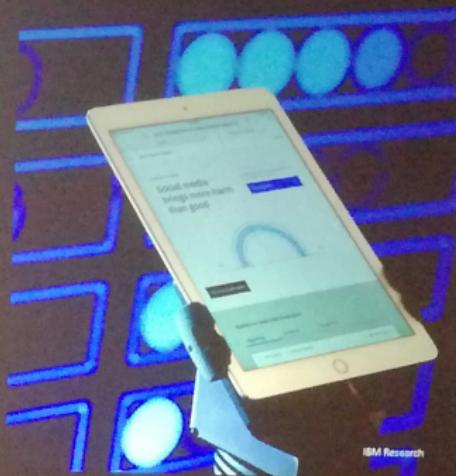
* Human opponent speech is recorded independently of the System's opening speech

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



Moving forward

- IBM Research is in a journey to develop technologies to master human language
- The Debater team mission is to develop language technologies to enhance decision-making in enterprises
- Informed decisions require considering pros and cons, typically done via Reading / Consulting
- Key example – Debater Speech by Crowd



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



Publications and Datasets are available at -



<https://www.research.ibm.com/artificial-intelligence/project-debater/research/>

21

IBM Research

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

@vivwyilai | @jon_tsiao_o | @chenhaotan
University of Colorado Boulder

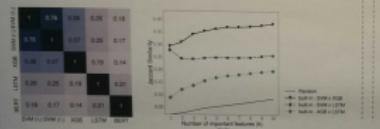


Example Yelp review: One of favorite places to eat on the King W side, simple and relatively quick. I typically always get the chicken burrito and the small is enough for me for dinner. Ingredients are always fresh and watch out for the hot sauce cause it's still scratching hot. Seating is limited so 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, <i>fresh</i> , quick, <i>favorite</i> the, dinner, be, quick, and, even, you, always, <i>fresh</i> , <i>favorite</i>	is, can, quick, <i>fresh</i> , at, to, always, even, <i>favorite</i> you, to, <i>fresh</i> , quick, at, can, even, always, and, <i>favorite</i>	me, be, relatively, enough, always, <i>fresh</i> , ingredients, prepared, quick, <i>favorite</i> dinner, ingredients, typically, <i>fresh</i> , places, cause, quick, and, <i>favorite</i> , always	.., ingredients, relatively, quick, places, enough, dinner, typically, me, i, one, watch, to, enough, limited, cause, and, <i>fresh</i> hot, <i>favorite</i>
LIME				

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

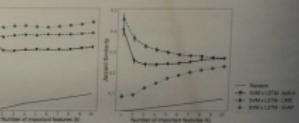


- Linear models have greatest similarity.
- LSTM with attention & BERT pay attention to different features.

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



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



- 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.

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

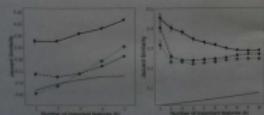
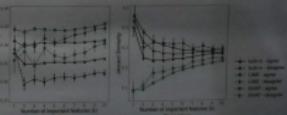


Figure: Similarity comparison between models with the built-in method.

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

✗ 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.
- Similarity is consistently higher when they agree than when they disagree.

EMNLP Posters



Visual Detection with Context for Document Layout Analysis

Carlos X. Soto and Shinjae Yoo

The Problem with PDFs

Automatic information extraction from PDFs (including scientific literature) is hindered by the following problems:

- Loss of page markup: preserved visual document appearance (e.g., for printing), but elements can be arbitrarily greater and superimposed, so source may be wildly inconsistent from document to document
- Result: standard text extraction tools for PDFs often yield noisy results
 - especially true for scientific articles (regularly edited and republished by publishers/document hosts, compounding markup issues)
 - text noise very inconsistent; cannot be cleaned up automatically; significantly hurts performance of downstream text analysis (e.g. NER)
- Also: information extraction from figures and tables remains very limited

Region-labeled Article Dataset

Was created to resolve issues of reading-region detection and context from the PMC Open Access subset. 9 regions selected were selected:

- Title: Includes subtitle, if present.
- Authors: Author names only where possible (i.e. no affiliations, etc.)
- Abstract: Abstract text only, where possible.
- Body: All main article text, including section headers. Contiguous, where possible.
- Figure: Any labeled figures (i.e. no figure logos, etc.)
- Figure Caption: The caption text for a figure.
- References: Full bibliography, not including post-references notes (e.g. author bio, journal marketing, etc.)
- Table: Including only the tabular contents, where possible. Includes adjacent notes or comments aligned.
- Table Caption: Main table name, as well as paragraph-table commentary that follows some tables.

Versions of this dataset is available at <https://github.com/cxsoto/article-detection>. It is in **PDF**, **XML**, and **JSON** formats, along with scripts to download and extract the original articles and to convert annotations to different formats.

Visual Document Segmentation with Context

Our approach is to ignore source markup and instead identify and segment salient document regions.

- Render articles as series of images, use object detection model to locate and classify regions
 - two-stage Faster R-CNN model as base model, uses pre-trained deep neural network (ResNet-101) to extract a feature map for input image
 - region proposal network generates candidate regions of interest (boxes containing likely "objects")
- Classification stage assigns labels to detected regions
- Requires training set of region-labeled articles
- Faster R-CNN chosen for ease of integrating contextual features

Unlike most images, scientific articles have unique features that we can exploit as contextual features of the classification and regression stages:

- The relative position and size of various regions within the document, as well as relative to the full article, are simple, but valuable contextual clues about the class of that region
- normalized and appended to the input features as input regions of interest prior to classification and bounding box regression
- Other contextual feature types are also being explored

Document Layout Detection Results

Adding even very simple contextual features led to >50% relative improvement in mean average precision.

- Also outperformed state of the art single-stage detectors (YOLOv3 and RetinaNet)
- Main body text can be accurately segmented with ~92% average precision
- Small regions (headers, titles, captions) require a much larger margin
- Working on larger training set, additional contextual features, and implementation in single-stage model

Mean Average Precision

Per-class Average Precision (Ours)



SIC

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

Lukas Lange^{1,2} Michael A. Hedderich² Dietrich Klakow²

{llange,mhedderich,dietrich.klakow}@sv.uni-saarland.de

¹Bosch Center for Artificial Intelligence, Renningen, Germany

²Spoken Language Systems (LSV), Saarland Informatics Campus,
Saarland University, Saarbrücken, Germany



BOSCH

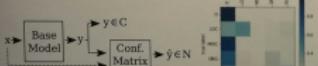
Invented for life

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 classifier's 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])



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

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

- ▶ Use pairs of clean and noisy labels for the same instances to initialize confusion matrix b_0
- ▶ 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 | Ez | NL |
|-----------|------|------|------|------|------|
| Precision | 23.2 | 39.9 | 51.0 | 59.7 | 23.4 |
| Recall | 9.2 | 30.1 | 24.7 | 49.3 | 21.1 |
| F1 | 15.2 | 34.3 | 33.3 | 54.0 | 25.6 |

Confusion Matrices for Different Word Cluster



Results

	De	En	Es	Ez	NL
Base (only clean data)	21.4 ± 1.0	36.9 ± 4.6	39.1 ± 3.6	36.7 ± 1.8	15.5 ± 2.7
Base (clean + noisy data)	26.2 ± 0.6	50.5 ± 1.4	50.2 ± 1.0	51.5 ± 0.7	39.5 ± 2.7
Global Noise Model	34.1 ± 1.4	52.0 ± 1.6	52.9 ± 0.6	52.3 ± 0.6	34.1 ± 2.8
(Velt 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 ± 5.5
Grouped Clusters	34.3 ± 1.4	51.4 ± 2.3	57.7 ± 2.4	53.1 ± 0.9	40.6 ± 1.3
K-Mean-Clauses	31.1 ± 2.1	57.6 ± 1.5	57.2 ± 1.3	55.2 ± 0.3	30.7 ± 1.0

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

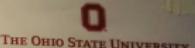
Conclusions

- ▶ 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%



LEVERAGING 2-hop DISTANT SUPERVISION FROM TABLE ENTITY PAIRS FOR RELATION EXTRACTION

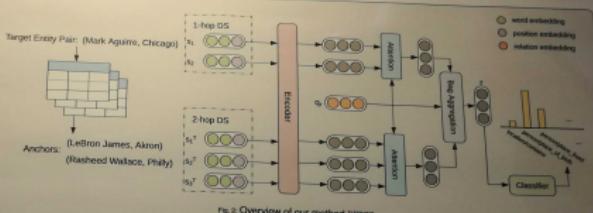
Xi Li, Ming Huan Sun
State University



Overview

Target Entity Pair	Mark Aguirre - Chicago place_head	Mark Aguirre and even monkeys , the mean presidents , few in charge to attend the funeral of my never .
Web Table	Mr. Basketball USA Westinghouse	Chicago, IL Silvers Gratz Westinghouse, PA
Anchor Entity Pairs	LeBron James St. Vincent - St. Mary	Akron, OH
	LeBron James - Akron	... including the eleven native Lebron James , given the dimensionless college town .
	Rasheed Wallace - Philly	another LeBron James , the high school player from Akron , today , who ... scored 13 points and Philadelphia native , Rashied Wallace , added 10 ...

Fig. 1: Illustration of 2-hop distant supervision.



Intuition

- There are massive Web tables that contain **relational facts** about entities.
- We can extract from them sets of entity pairs that share common relations, which we refer as **anchors** for each other.
- We can use anchor entity pairs of a given target entity pair to help infer the relation, which we refer as **2-hop distant supervision**.

Contributions

- We introduce **2-hop DS** as an extension to the conventional distant supervision, and leverage entity pairs in Web tables as anchors to **find additional supporting sentences** to further improve RE.
- We propose **REDS2**, a new neural relation extraction method based on 2-hop DS and has achieved new state-of-the-art performance in the benchmark dataset.
- We release both our **source code** and an **augmented benchmark dataset** that has entity pairs aligned with those in Web tables, to facilitate future work.

Problem Formulation

Finally, we define entity pairs that potentially have the same relation with a target entity pair as anchors, which can be found through Web tables, to fully exploit the sentences that mention those anchor entity pairs to support RE for the target entity pair.

Experiments

# Entity Pairs	Train Overall	Test Overall	Train Non-NA	Test Non-NA
291699	18144	95678	1761	

Data&Code





NET^x ++ Low-Resource Name Tagging Learned with Weakly Labeled Data

Yixin Cao, Zikun Hu, Tat-Seng Chua, Zhiyuan Liu, Heng Ji
 caoyixin2011@gmail.com, zikunhu@u.nus.edu, dcsts@nus.edu.sg
 liuzy@tsinghua.edu.cn, hengji@illinois.edu

Background

- Name tagging in low-resource languages or domains suffers from inadequate training data.
- Existing works focus on transferring extra knowledge which are effective only in specific languages or domains.
- Weakly Labeled (WL) data is cheap to collect and not explored yet. But to utilize it in name tagging is challenging because it is (1) partially labeled; (2) noisy with missing labels and incorrect annotation.

Fully Labelled, expensive to obtain

s_1 : [B-ORG L-ORG 0 0 0 0 B-LOC]
 Formula shell won game one in Philippines ...

s_2 : [B-MT L-MT B-ORG L-ORG]
 Barangay Ginebra and Formula Shell forming a rivalry ...

Weakly Labelled, extensively exists on the web

Method

- Data selection scheme
- Divide noisy dataset into high-quality data and noisy data from two aspects.

$$g(X, Y) = \frac{\sum_{(x_i, y_i) \in D} I(y_i \in \mathcal{Y}) p(x_i|C(y_i)) p(C(y_i|x_i))}{|D|}$$

$$n(X, Y) = \frac{\sum_{(x_i, y_i) \in D} I(y_i \in \mathcal{Y})}{|D|}$$

- Pre-train classification module on noisy data
- Loss function:

$$\hat{L}_c = -\sum_{(x, y)} I(y \in \mathcal{Y}) \hat{y}_i \log P_{c,y_i}$$

- Non entity sampling
- Since no word is labeled with O in WL data, we sample O from unlabeled words.

$$p(y_i = O | x_i, y_i = N) = \frac{1}{2} (\lambda_0 f_1 + \lambda_2 (1 - f_1) + \lambda_4 f_3)$$

$p(\hat{Y}|X) = \prod_{(X, Y)} p(Y|X)$

Loss function:

$$\mathcal{L} = -\sum_{(X, \hat{Y})} \log p(\hat{Y}|X)$$



A Bayesian Approach for Sequence Tagging With Crowds

Edwin Simpson, Iryna Gurevych
Ubiquitous Knowledge Processing Lab, Technische Universität Darmstadt

Technische Universität Darmstadt

Experiments

Datasets

- NER [5]: CoNLL 2003 named entities, 4 classes, 5 annotators/doc
- PICO [4]: spans identify medical trial populations in biomedical literature, 6 annotators/doc, mid-size spans (avg. 8 tokens)
- ARG [6]: pro and con argument spans, 5 annotators/doc, long spans (avg. 18 tokens)

How to combine sequence labels from multiple, unreliable sources?

Core idea

- Weighting annotators > majority voting
- But many annotators label small amounts of data, → no aggregation models (weights) tend to overfit
- So use a Bayesian approach to account for model uncertainty

From [7], [8]

Most previous methods do not handle sequential data

- David-Skene [5]: probabilistic confusion matrices
- IBCC [2]: adds priors + Bayesian inference
- MACE [3]: models spamming patterns
- HMM+crowd [4]: models dependencies between true labels only

The BSC model

- The true label sequence is modelled by an HMM, which defines the probability of one label following another
- Given the true label, the HMM observation model defines the likelihood of a token in the vocabulary and the annotator model defines the likelihood of the label from each annotator
- Learns the complete model in an unsupervised or semi-supervised way (with some gold labels) using variational Bayesian inference

Annotator Models

IBCC can use any of these annotator models:

- Spammer model (spam): E.g. MACE vs. IBCC
- Aggressive

F1 scores

	NER	PICO	ARG
Best worker	67.3	58.5	60.0
Majority vote	65.4	64.3	34.8
MACE	70.0	39.0	32.0
Dawid-Skene	74.4	68.7	47.4
IBCC	74.4	68.0	47.4
HMM	74.4	68.0	47.4

E.g.: on the topic of legalising marijuana use, 2 workers found a span:
The dangers of marijuana use have been exaggerated for long and...

E.g.: on the topic of stem cell research, 3 workers found a span:
oC Stem cells can be cloned to deal with cardiac failures this



Learning Entity Representations

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.
- Entity representations (DER) that rely on entries in Wikipedia.

EntEval

7 probing task groups.

Entity Typing (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 ...

Reference Arc Prediction (CAP)
CAP = classify if two entities are the same given context.

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

Factuality Prediction (EFP)
EFP = classify the correctness of statements for entities.

TD Garden has held Bruins games. ✓

Contextualized Entity Relationship Prediction (CER)
CER = 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.

SOCCKER - JAPAN GET LUCKY WIN, CHINA IN SURPRISE

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	Epochs	Train	CAP	CP	EFP	ET	NED	ESR	EFT
NED	CONLL-YAGO	≤ 30	epochless	2	2	2	10301	NA	626	

Dataset References

- ET: Ultra-scale entity typing
- CAP: CoNLL-X: A large-scale dataset in preschool vocabulary for coref resolution.
- CP: Concrenet 5.5: An open multilingual graph of general knowledge.
- NED: Robust disambiguation of named entities in text.
- EFP: Rare entity prediction with hierarchical items under external descriptions.
- ET: Kore: Keyphrase extraction from news articles for distant supervision.
- ESR: Jointly learning entities and text with distant supervision.
- EFT: Jointly learning entities and text with distant supervision.

EntEval: A Holistic Evaluation Benchmark for Entity Representations

Mingda Chen^{*1}, Zewei Chu^{*2}, Yang Chen⁴, Karl Stratos³, Kevin Gimpel¹

¹Toysota Technological Institute at Chicago ²University of Chicago ³Rutgers University ⁴Ohio State University

*Equal Contribution. Listed in alphabetical order.

Hyperlink-Based Training

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

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

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

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

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

Special symbols prepended to sentences to indicate the start of descriptions from contexts.

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

The final training loss for EntELMo is

$$l_{lang}(x_{1:T_s}) + l_{lang}(y_{1:T_d}) + l_{ELM} + l_{desc}$$

Experiment Results

	ET	CAP	EFP	NED	CP	ERT	ESR
GloVe	10.3	71.9	67.0	41.2	32.8	40.8	50.9
BERT Base	32.0	80.6	74.8	50.8	65.8	42.2	28.8
BERT Large	32.3	79.1	76.7	54.3	66.9	48.6	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_{desc}	33.2	73.5	71.1	48.9	59.4	44.6	53.3
EntELMo w/ l_{desc}	33.6	76.3	70.9	49.3	60.4	42.9	60.3

Table 2. EntELMo w/ l_{desc} is trained with a modified version of l_{desc} 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
Ganes and Hofmann, 2017	66.7

Scan to check out the code and data



ILLINOIS

TIGER: Text-to-Image Grounding for Image Caption Evaluation

Ming Jiang¹, Qiuyuan Huang², Lei Zhang², Xin Wang³, Pengchuan Zhang², Zhe Gan², Jana Diesner¹, Jianfeng Gao²
Microsoft Research
UCSB

1. University of Illinois at Urbana-Champaign
2. Microsoft Research
3. University of California Santa Barbara

Motivation and Contribution

Existing metrics based on text-level comparisons do not consider image information and do not address language ambiguity.

- Propose a novel automatic evaluation metric called TIGER.
- Consider both image content and human-generated references.
- Measure the consistency of human attention distribution across image regions.

TIGER Framework

Data Encoding

- Region-level & Word-level embedding vectors

Text-to-Image Grounding

- Grounding a caption into each image region.
- [Reference vs. Candidate] | Image
- RRS: how well is the order of image regions matched with descriptions based on grounding weights?
- WDS: how similar is the attention distributed to the image by a caption across image regions?

TIGER

- Average value of RRS and WDS

TIGER Workflow

- Encoding images and texts by a pre-trained Bottom-Up Attention and a RNN model.
- Grounding texts and images by a pre-trained SCAN model.
- Calculating RRS based on Normalized Discounted Cumulative Gain (NDCG).
- Measuring WDS based on KL-Divergence.

Metric Performance

- TIGER achieved a noticeable improvement in assessing caption quality on three benchmark datasets.
- Identifying relevant human-written captions in HC is relatively easy for all metrics, while judging the quality of two correct human-annotated captions in HC is more with metrics than other competition groups.
- By changing the reference sizes, TIGER achieves a higher judgment accuracy and more stable performance.

Result Analysis

- Image region has a higher grounding weight with the corresponding caption than other unrelated regions.
- Text-to-image grounding is more difficult at action-level compared to object-level.
- Reference captions may not fully cover visual information and TIGER can measure a caption quality by considering the semantic information of image contents.
- Human interpretation inspired by the image is hard to be judged by an automatic evaluation metric.

	Composite	Human
r	0.388	0.479
p	0.023	0.031
AUC	0.205	0.263
BLEU-1	0.397	0.468
BLEU-4	0.265	0.338
METEOR	0.397	0.468
CIDEr	0.318	0.418
LCS	0.318	0.418
Ours	0.388	0.479
Spearman	0.437	0.508
TIGER	0.454	0.523

Caption-based correlation between metrics and ground truth. Human scores are generated by using Kendall tau and Spearman rho (Hoeffding's D) test ($p < 0.05$).

Accuracy of various in matching human judgments on TIGER-10k with 1 reference caption. The highest accuracy is achieved by TIGER, followed by SQuAD, METEOR, and CIDEr. Human scores are generated by using Kendall tau and Spearman rho (Hoeffding's D) test ($p < 0.05$).

Qualitative Results

Human: 1 A curly-haired child is blowing away dandelion seeds while laying in a field of lush green grass.

Human: 2 As he lays in the grass, he picks a dandelion seed and blows it away.

TIGER: 1 A boy laying in the grass, the boy is blowing a dandelion seed.

Human: 3 A professional baseball game being played at night.

TIGER: 4 A baseball player getting ready to throw the ball from a base.

Interpretation Metrics (IM) Plots

IM Plots show the distribution of grounding weights for different regions. The plots include histograms and cumulative distribution functions (CDFs) for various regions (e.g., object, property, action).

Qualitative Results

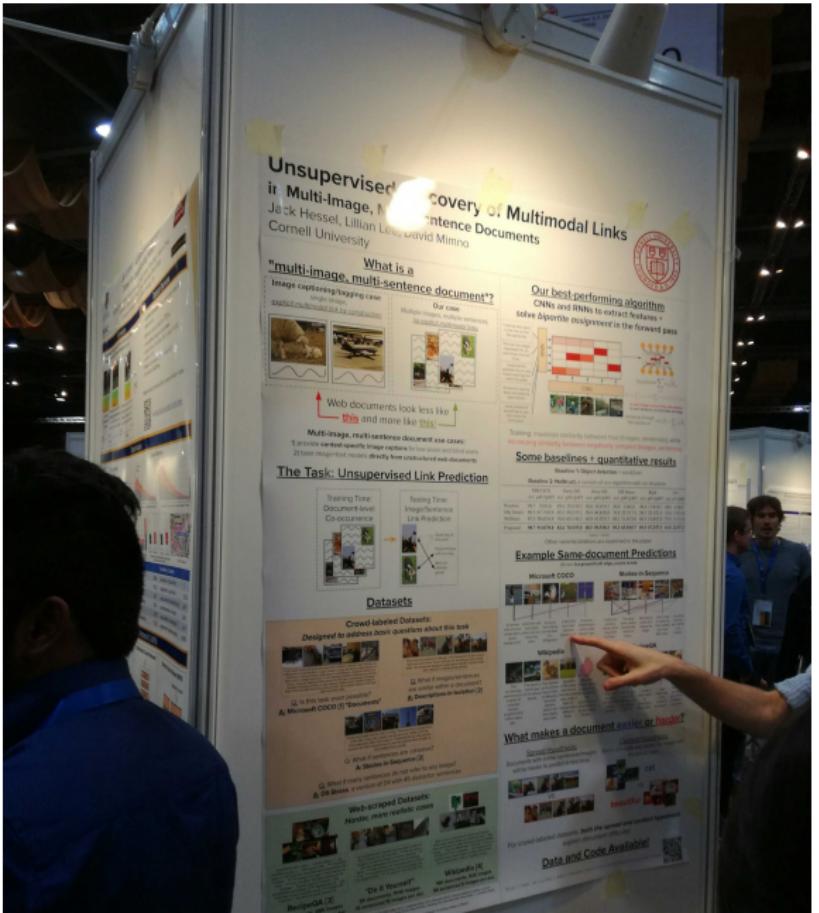
A bunch of people are in an open court yard.

A young kid hitting a baseball with a bat close to his head.

Interpretation Metrics (IM) Plots

IM Plots show the distribution of grounding weights for different regions. The plots include histograms and cumulative distribution functions (CDFs) for various regions (e.g., object, property, action).

EMNLP Posters





Hierarchical Meta-Embeddings for Code-Switching Named Entity Recognition

Genta Indra Winata¹, Zhaqiang Lin¹, Jian Liu², Zihan Liu¹, Pascale Fung¹

¹Center for Artificial Intelligence Research (CAIRE), The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
²<https://github.com/gentaiscool/meta-emb>

Background
 "walking dead is quite of a aperto a cquinquena"
 (single TV news site)
 (multiple news sites from different countries)

- Previous work mainly focused on word-level aspects, however, languages share common subwords especially for closely related languages and/or for languages that are severely imbalanced.
- We propose Hierarchical Meta-Embeddings (HME) that learn word, subword and character level embeddings to create language-agnostic lexical representations.
- We also show that, in cross-lingual settings, our model not only leverages closely related languages, but also learns from languages with different roots.

Methodology

Multilingual Meta-Embedding (MMR)

- We generate meta-representations by taking the linear representation of three multiple monolingual pre-trained embeddings.

- We apply a projection matrix \mathbf{W}_t to transform the dimension from the original space $\mathbf{x}_{i,t} \in \mathbb{R}^d$ to a new shared space $\mathbf{x}'_{i,t} \in \mathbb{R}^d$. Then, we calculate attention weights $a_{i,t}$ in \mathbb{R}^d with a non-linear scoring function.

- Then, MMR is calculated by taking the weighted sum of the projected embeddings $\mathbf{x}'_{i,t}$

$$\mathbf{x}'_{i,t} = \mathbf{W}_t \cdot \frac{\exp(\phi_i(\mathbf{x}_{i,t}))}{\sum_j \exp(\phi_j(\mathbf{x}_{j,t}))} \quad (1)$$

$$a_{i,t} = \frac{\exp(\phi_i(\mathbf{x}'_{i,t}))}{\sum_j \exp(\phi_j(\mathbf{x}'_{j,t}))} \quad (2)$$

$$\mathbf{u}_i = \sum_t a_{i,t} \phi_i(\mathbf{x}_{i,t}) \quad (3)$$

Mapping Subwords and Characters into Word-level Representations

- We map each segment words into sets of BPEs, and then we extract the most frequent subword embedding vectors $\mathbf{x}^{subword}_j \in \mathbb{R}^d$ for each word $j \in \mathcal{W}$. We replace the projection matrix with Transpose. Then, we create $\mathbf{u}^{subword}_i \in \mathbb{R}^d$ which represents the subword-level HME by taking the weighted sum of $\mathbf{x}^{subword}_j \in \mathbb{R}^d$.

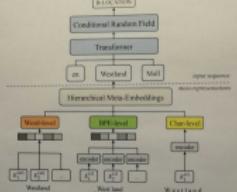
$$\mathbf{x}^{subword}_j = \text{Encoder}(\mathbf{x}^{subword}_j) \quad (4)$$

$$\mathbf{u}^{subword}_i = \sum_j a_{j,i} \mathbf{x}^{subword}_j \quad (5)$$

- We combine character-level representations, we apply an encoder to each character

Hierarchical Meta-Embeddings (HME)

- We confuse word, subword, and character representations to create a mixture of embeddings. Let \mathbf{w} be a sequence of words with k elements, where $\mathbf{w} = [\mathbf{w}_1, \dots, \mathbf{w}_k]$.
- Each word can be tokenized into a list of subwords $\mathbf{s} = [s_1, \dots, s_n]$ and a list of characters $\mathbf{c} = [c_1, \dots, c_p]$. The list of subwords \mathbf{s} is generated using a function $f_s: \mathbf{w} \rightarrow \mathbf{s}$. Function f maps a word into a sequence of subwords.
- Further, let E^w , E^{sub} and E^c be a set of word, subword, and character embedding lookup tables.



Experimental Results

Model	Multilingual embeddings		Cross-lingual embeddings		Model	Cross-lingual embeddings	
	Word	Char	Word	Char		Word	Char
<i>Plain word embeddings</i>							
DINCFAT	40.3 ± 0.0	60.39 ± 0.0	65.01 ± 1.0	65.01 ± 1.0	Word	60.20 ± 0.0	64.03 ± 0.0
LINFORA	54.41 ± 0.0	61.21 ± 0.0	65.48 ± 0.0	65.48 ± 0.0	Char	50.77 ± 1.0	50.77 ± 1.0
Word	50.43 ± 0.0	60.00 ± 0.0	65.00 ± 0.0	65.00 ± 0.0	Our approach	60.80 ± 0.0	63.20 ± 0.0
Subword	50.43 ± 0.0	60.00 ± 0.0	65.00 ± 0.0	65.00 ± 0.0	Essential	60.80 ± 0.0	63.40 ± 0.0
<i>Hierarchical Meta-Embedding (HME)</i>							
Char	50.86 ± 0.0	61.70 ± 0.0	65.00 ± 0.0	65.00 ± 0.0	Our approach	67.3 ± 0.0	69.17
Subword	50.86 ± 0.0	61.70 ± 0.0	65.00 ± 0.0	65.00 ± 0.0	Essential	67.25 ± 0.0	69.17
Word	50.86 ± 0.0	61.70 ± 0.0	65.00 ± 0.0	65.00 ± 0.0	Our approach	68.53 ± 0.0	69.99

Analysis

- Word-level meta-embeddings without subword or character-level information consistently perform better than flat baseline settings.
- Adding subword inputs to the model is consistently better than character. Subword embeddings are more effective for distant languages than closely related languages.
- The model usually chooses the correct language embeddings.

Conclusion

- HME combines multiple monolingual word-level and subword-level embeddings to create language-agnostic representations without any specific language information.
- HME leverages subword information very effectively from languages that are severely imbalanced.



EMNLP-JNLPB 2019

Improved Differentiable Architecture Search for Language Modeling and Named Entity Recognition

Yufan Jiang¹, Chi Hu¹, Tong Xiao^{1,2}, Chunliang Zhang^{1,2}, Jingbo Zhu^{1,2}

NLP Lab, Northeastern University¹

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.

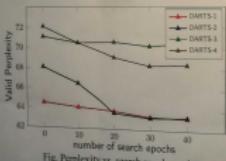
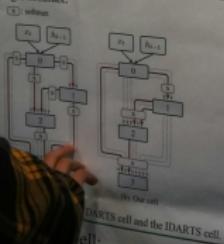


Fig. Perplexity vs. search epoch number.

Improved DARTS (IDARTS)

- ❑ 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

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.

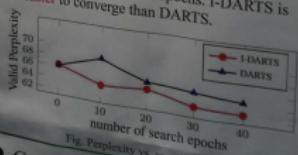
Architecture	Perplexity		Search Cost (GPU days)
	val	test	
V-RNN	67.9	65.4	-
LSTM	60.7	58.8	-
LSTM + SC	60.9	58.3	-
LSTM + SE	58.1	56.0	-
ENAS	60.8	58.6	0.50
DARTS	58.3	56.1	0.25
Random RNNs	63.7	61.2	-
I-DARTS ($n = 1$)	58.0	56.0	0.17
I-DARTS ($n = 2$)	-	-	-

Table. Perplexities on PTB (lower is better)

Model	F1
<i>best published</i>	
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-M-CRF+PCF (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.



Conclusion

- ❑ Im-

Connections



Thank you!