Leveraging External Knowledge

On different tasks and various domains

Gabi Stanovsky

(a somewhat obvious) Introduction

- Performance relies on the amount of training data
- It is expensive to get annotated data on a large scale
- Can we use external knowledge as additional signal?

In this talk

- Recognizing adverse drug reactions in social media
  - Integrating knowledge graph embeddings
- Factuality detection
  - Using multiple annotated datasets
- Acquiring predicate paraphrases
  - Using Twitter metadata and syntactic information

Recognizing Mentions of Adverse Drug Reaction

Gabriel Stanovsky, Daniel Gruhl, Pablo N. Mendes

EACL 2017

Recognizing Mentions of Adverse Drug Reaction in Social Media

Gabriel Stanovsky, Daniel Gruhl, Pablo N. Mendes

Bar-Ilan University, IBM Research, Lattice Data Inc.

April 2017

In this talk

1. **Problem:** Identifying adverse drug reactions in social media
   - "*I stopped taking* ***Ambien*** *after three weeks, it gave me a* ***terrible headache***"

In this talk

1. **Problem:** Identifying adverse drug reactions in social media
   - "*I stopped taking* ***Ambien*** *after three weeks, it gave me a* ***terrible headache***"
2. **Approach**
   - LSTM transducer for BIO tagging
   - + Signal from knowledge graph embeddings

In this talk

1. **Problem:** Identifying adverse drug reactions in social media
   - "*I stopped taking* ***Ambien*** *after three weeks, it gave me a* ***terrible headache***"
2. **Approach**
   - LSTM transducer for BIO tagging
   - + Signal from knowledge graph embeddings
3. **Active learning**
   - Simulates a low resource scenario

Task Definition

Adverse Drug Reaction (ADR)

Unwanted reaction clearly associated with the intake of a drug

- We focus on automatic ADR identification on social media

Motivation - ADR on Social Media

1. Associate unknown side-effects with a given drug
2. Monitor drug reactions over time
3. Respond to patients' complaints

CADEC Corpus (Karimi et al., 2015)

ADR annotation in forum posts (*Ask-A-Patient*)

- Train: **5723 sentences**
- Test: **1874 sentences**

Drug Ratings for AMBIEN

Average Rating: 3.2 (1408 Ratings)

RATING	REASON	SIDE EFFECTS FOR AMBIEN
1	insomnia due to MS	Sleep was disturbed by waking and vivid dreams. Day after side effects are horrible- dizziness, nausea, diarrhea, headache, severe depression.
1	insomnia	Woke up off and on all night headaches vivid disturbing dreams, heightened senses too much so change in mood aggressiveness

Challenges

Challenges

- **Context dependent**
  "*Ambien gave me a* ***terrible headache***"
  "*Ambien made my* ***headache*** *go away*"

Challenges

- **Context dependent**
  "*Ambien gave me a* ***terrible headache***"
  "*Ambien made my* ***headache*** *go away*"

- **Colloquial**
  "*hard time getting some Z's*"

Challenges

- **Context dependent**
  "*Ambien gave me a* ***terrible headache***"
  "*Ambien made my* ***headache*** *go away*"
- **Colloquial**
  "*hard time getting some Z's*"
- **Non-grammatical**
  "*Short term more loss*"

Challenges

- **Context dependent**
  "*Ambien gave me a* ***terrible headache***"
  "*Ambien made my* ***headache*** *go away*"
- **Colloquial**
  "*hard time getting some Z's*"
- **Non-grammatical**
  "*Short term more loss*"
- **Coordination**
  "*abdominal gas, cramps and pain*"

Approach:
LSTM with knowledge graph embeddings

Task Formulation

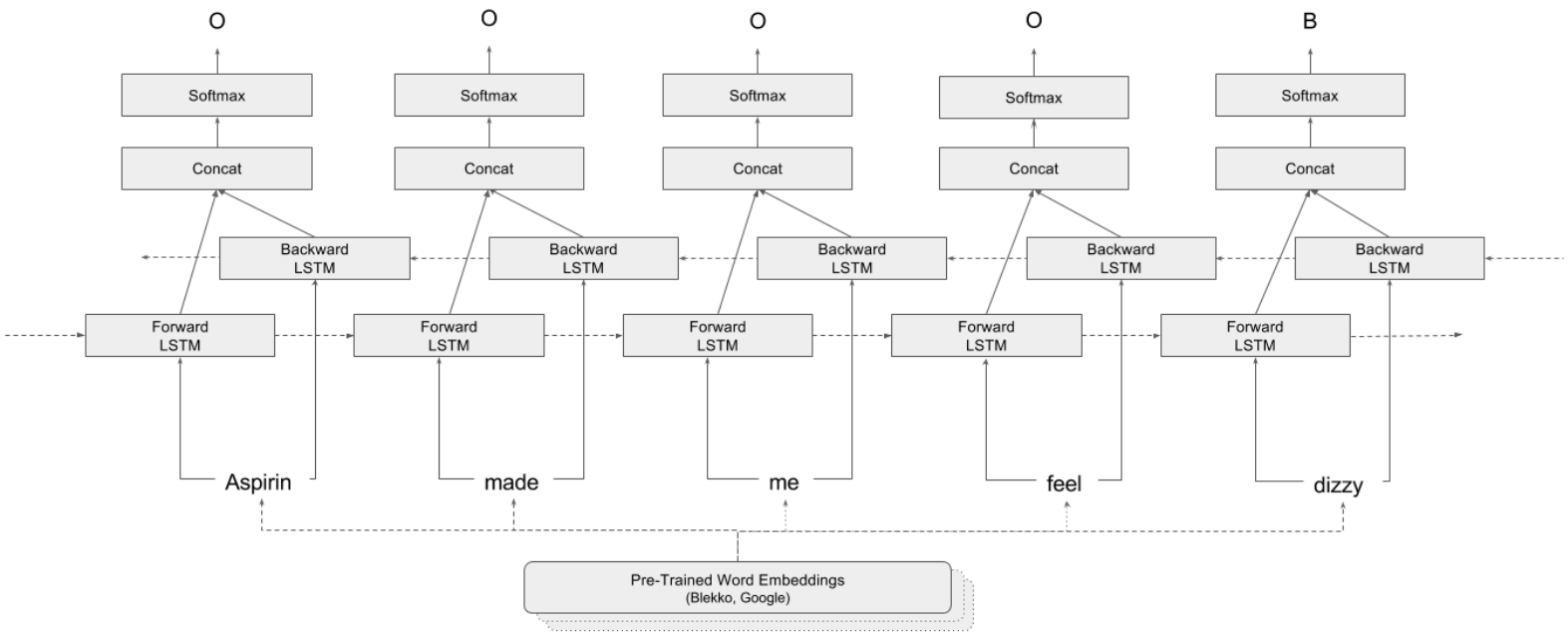
Assign a Beginning, Inside, or Outside label for each word

Example

"$[I]_O$ $[stopped]_O$ $[taking]_O$ $[Ambien]_O$ $[after]_O$ $[three]_O$ $[weeks]_O$ – $[it]_O$ $[gave]_O$ $[me]_O$ $[a]_O$ $[\textbf{terrible}]_{ADR\text{-}B}$ $[\textbf{headache}]_{ADR\text{-}I}$"

Model

- bi-RNN transducer model
  - Outputs a BIO tag for each word
  - Takes into account context from both past and future words



Integrating External Knowledge

- DBPedia: Knowledge graph based on Wikipedia
  - (**Ambien**, *type*, **Drug**)
  - (**Ambien**, *contains*, **hydroxypropyl**)

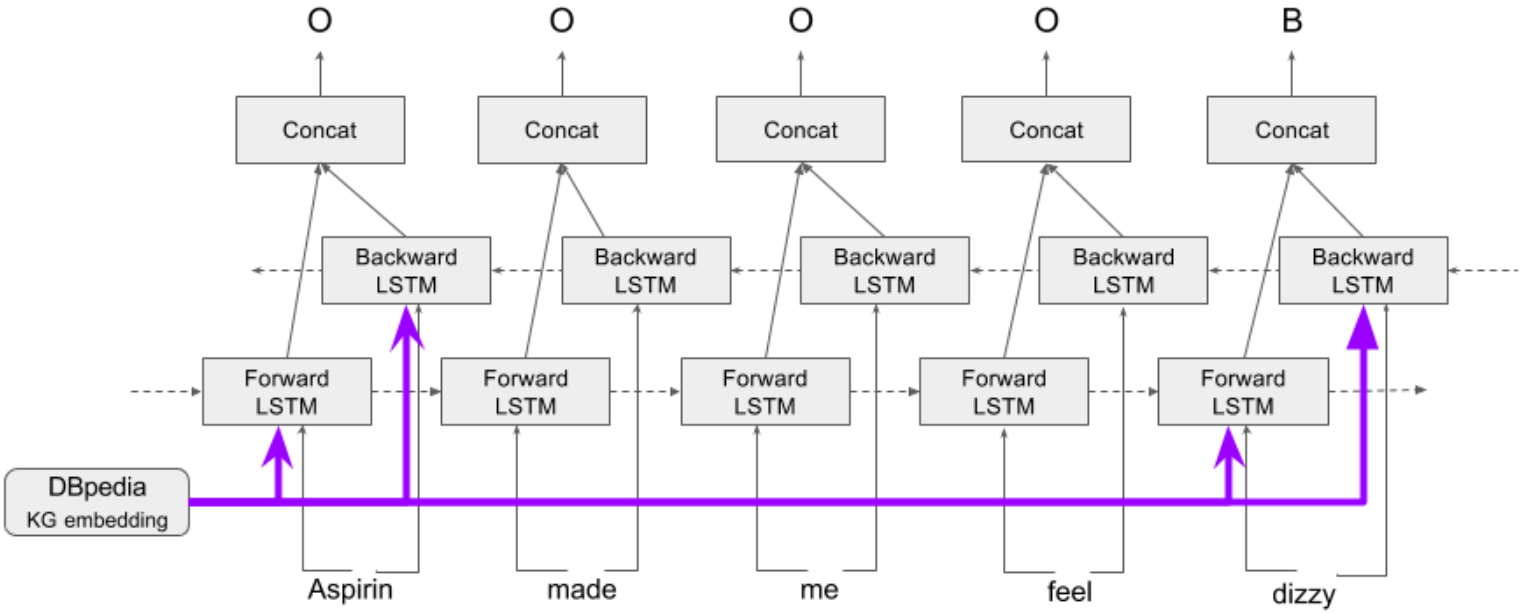
Integrating External Knowledge

- DBPedia: Knowledge graph based on Wikipedia
  - (**Ambien**, *type*, **Drug**)
  - (**Ambien**, *contains*, **hydroxypropyl**)
- Knowledge graph embedding
  - Dense representation of entities
  - Desirably:
    Related entities in DBPedia $\iff$ Closer in KB-embedding

Integrating External Knowledge

- DBPedia: Knowledge graph based on Wikipedia
  - (**Ambien**, *type*, **Drug**)
  - (**Ambien**, *contains*, **hydroxypropyl**)
- Knowledge graph embedding
  - Dense representation of entities
  - Desirably:
    Related entities in DBPedia $\iff$ Closer in KB-embedding
- We experiment with a simple approach:
  - Add verbatim *concept* embeddings to word feats

Prediction Example



Evaluation

	P	R	F1
ADR Oracle	55.2	100	71.1

- ADR Orcale - Marks gold ADR's regardless of context
  - Context matters $\rightarrow$ Oracle errs on 45% of cases

Evaluation

	Emb.	% OOV	P	R	F1
ADR Oracle			55.2	100	71.1
LSTM	Random		69.6	74.6	71.9
LSTM	Google	12.5	85.3	86.2	85.7
LSTM	Blekko	7.0	90.5	90.1	90.3

- ADR Orcale - Marks gold ADR's regardless of context
  - Context matters $\rightarrow$ Oracle errs on 45% of cases
- External knowledge improves performance:
  - Blekko $>$ Google $>$ Random Init.

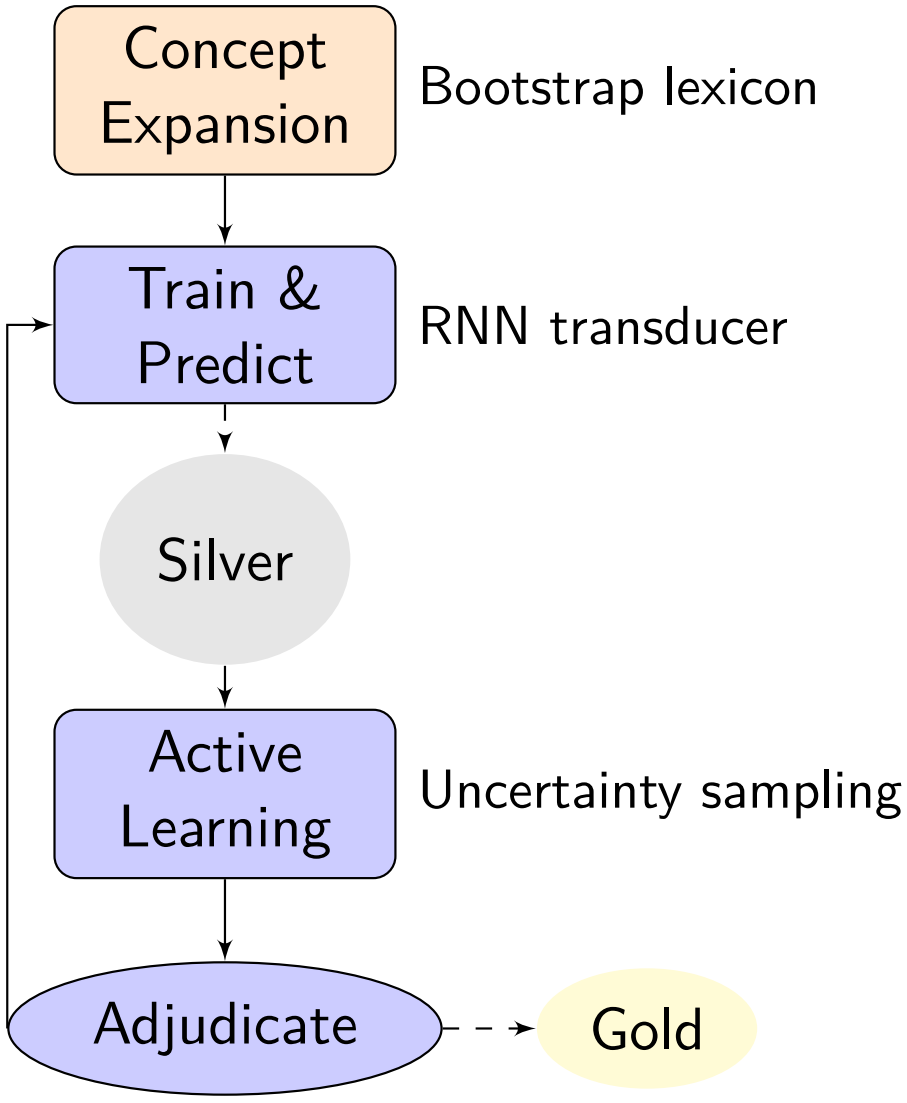
Evaluation

	Emb.	% OOV	P	R	F1
ADR Oracle			55.2	100	71.1
LSTM	Random		69.6	74.6	71.9
LSTM	Google	12.5	85.3	86.2	85.7
LSTM	Blekko	7.0	90.5	90.1	90.3
LSTM + DBPedia	**Blekko**	**7.0**	**92.2**	**94.5**	**93.4**

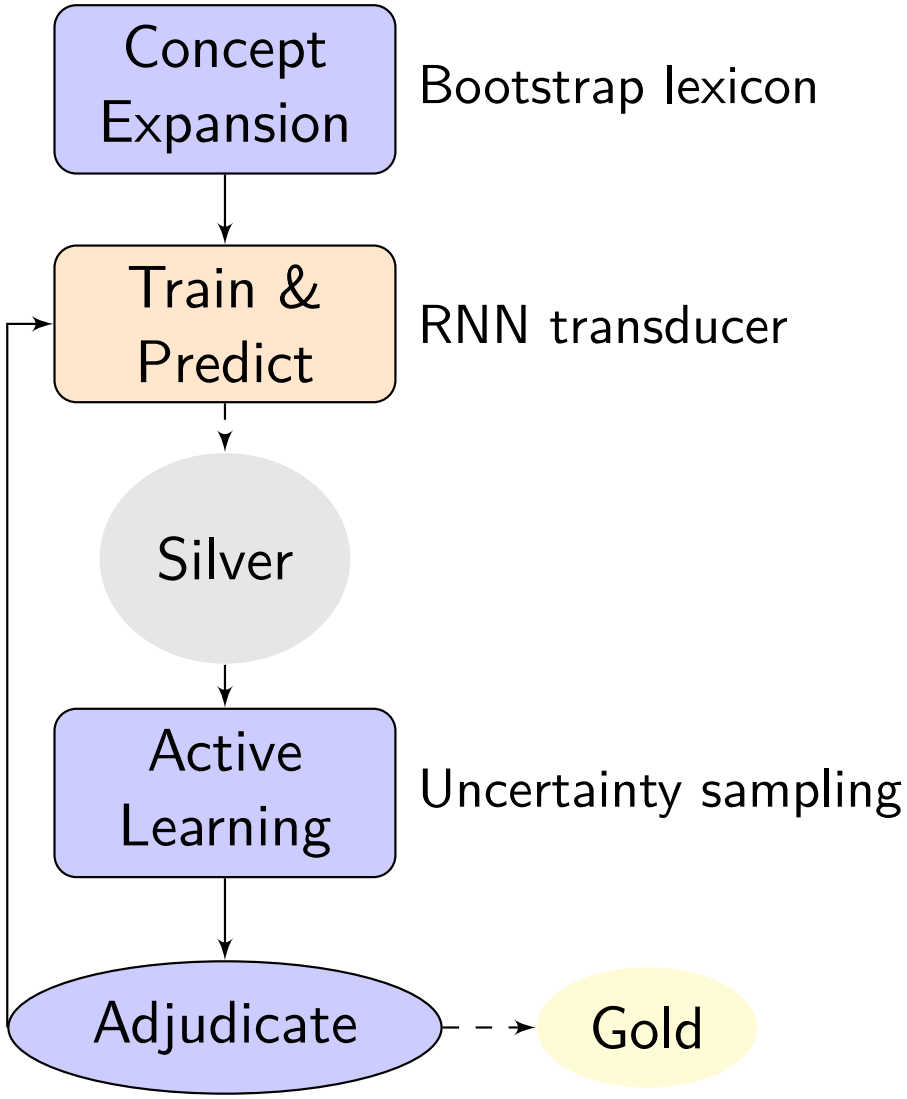
- ADR Orcale - Marks gold ADR's regardless of context
  - Context matters $\rightarrow$ Oracle errs on 45% of cases
- External knowledge improves performance:
  - Blekko $>$ Google $>$ Random Init.
  - DBPedia provides embeddings for 232 (4%) of the words

Active Learning:
Concept identification for low-resource tasks

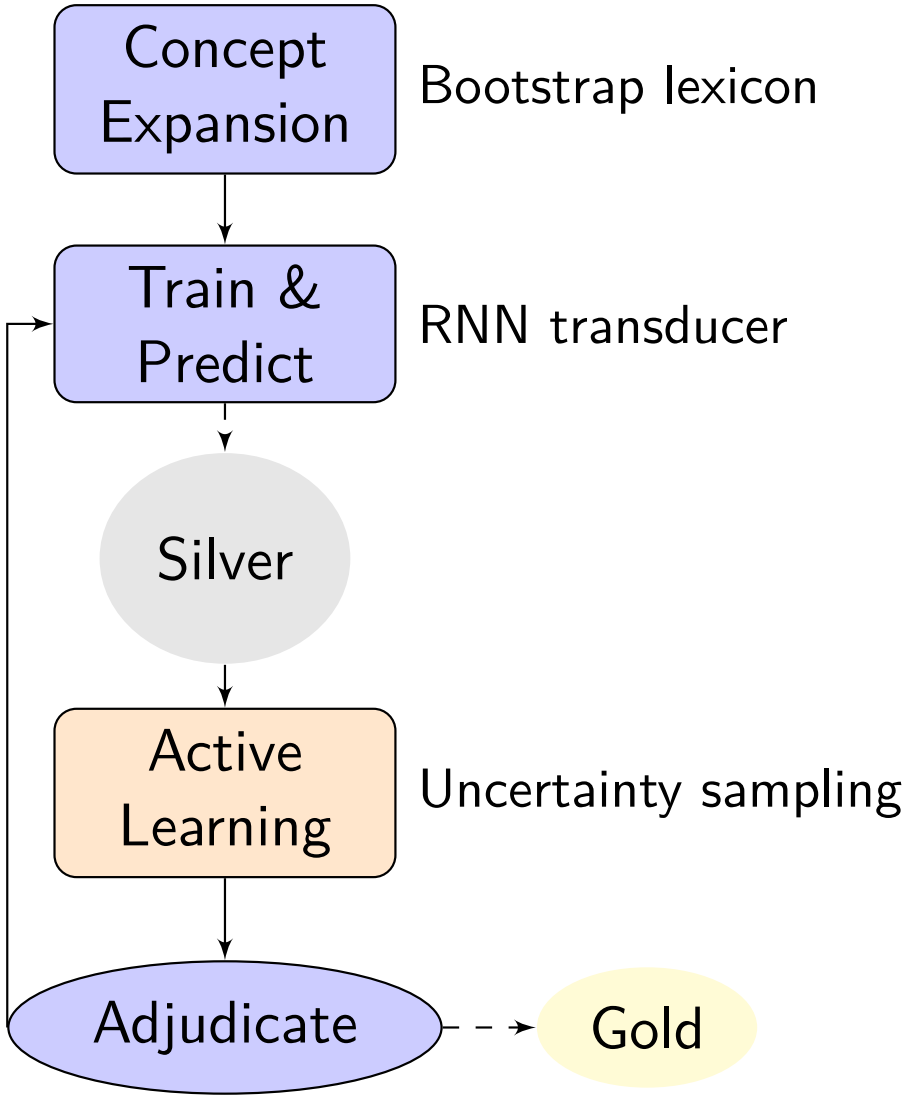
Annotation Flow



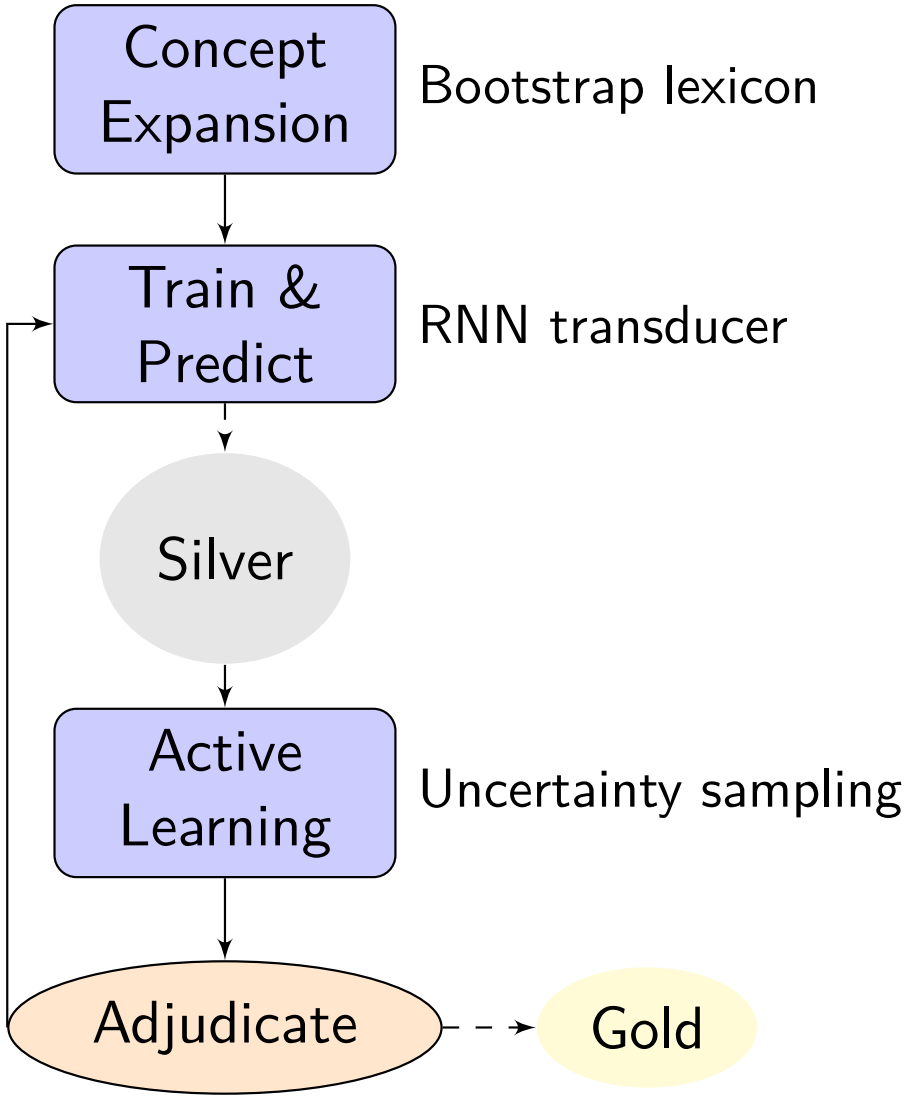
Annotation Flow



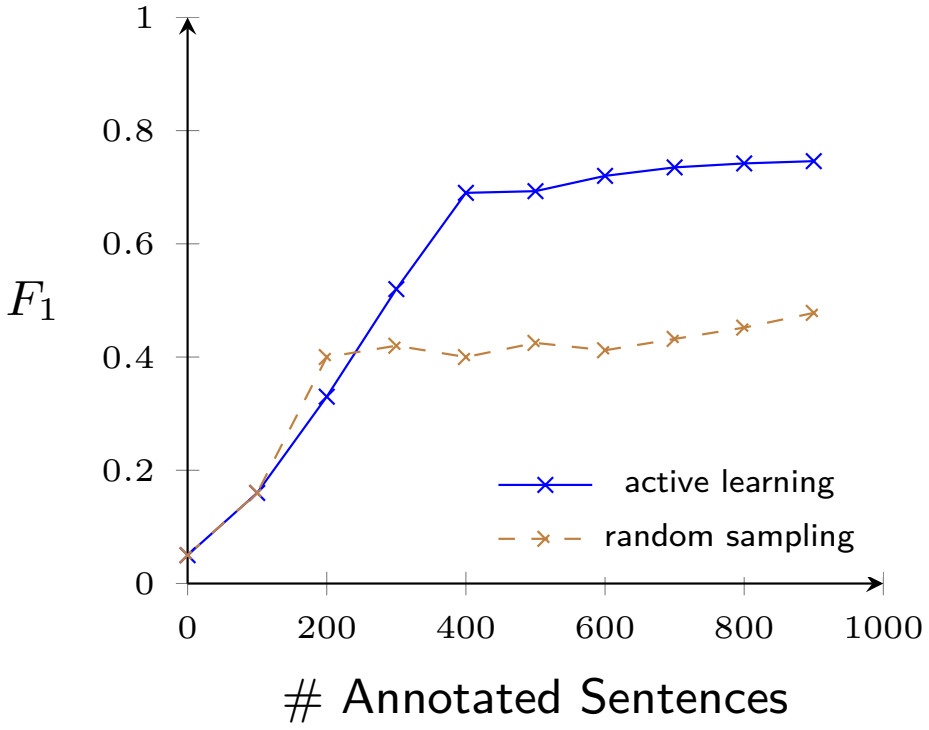
Annotation Flow



Annotation Flow



Training from Rascal



- Performance after 1hr annotation: **74.2 F1 (88.8 P, 63.8 R)**
- Uncertainty sampling boosts improvement rate

Wrap-Up

Future Work

- Use more annotations from CADEC
  - E.g., symptoms and drugs
- Use coreference / entity linking to find DBPedia concepts

Conclusions

- LSTMs can predict ADR on social media
- Novel use of knowledge base embeddings with LSTMs
- Active learning can help ADR identification in low-resource domains

Conclusions

- LSTMs can predict ADR on social media
- Novel use of knowledge base embeddings with LSTMs
- Active learning can help ADR identification in low-resource domains

Thanks for listening!
Questions?

Factuality Prediction over Unified Datasets

Gabriel Stanovsky, Judith Eckle-Kohler, Yevgeniy Puzikov,

Ido Dagan and Iryna Gurevych

ACL 2017

Outline

- Factuality detection is a difficult semantic task
  - Useful for downstream applications
- Previous work focused on *specific* flavors of factuality
  - Hard to compare results
  - Hard to port improvements
- We build a unified dataset and a new predictor
  - Normalizing annotations
  - Improving performance across datasets

Factuality

Task Definition

- Determining author's commitment
  - It is *not surprising* that **the Cavaliers lost the championship**
  - She still *has to check* whether **the experiment succeeded**
  - Don was *dishonest* when **he said he paid his taxes**

- Useful for
  - Knowledge base population
  - Question answering
  - Recognizing textual entailment

Annotation

- **Many shades of factuality**
  - She *might* sign the contract
  - She will *probably* get the grant
  - She *should not* accept the offer
  - ….

- A continuous scale from factual to counter-factual
  (Saur´ı and Pustejovsky, 2009)

Datasets

Corpus	#Tokens/Sentences	Factuality Values	Type	Annotators	Perspective
FactBank	77231 / 3839	Factual (CT+/-) Probable (PR+/-) Possible (PS+/-) Unknown (Uu/CTu)	Discrete	Experts	Author's and discourse-internal sources
MEANTIME†	9743 / 631	Fact / Counterfact Possibility (uncertain) Possibility (future)	Discrete	Experts	Author's
UW	106371 / 4234	[-3.0, 3.0]	Continuous	Crowdsource	Author's

- Datasets differ in various aspects

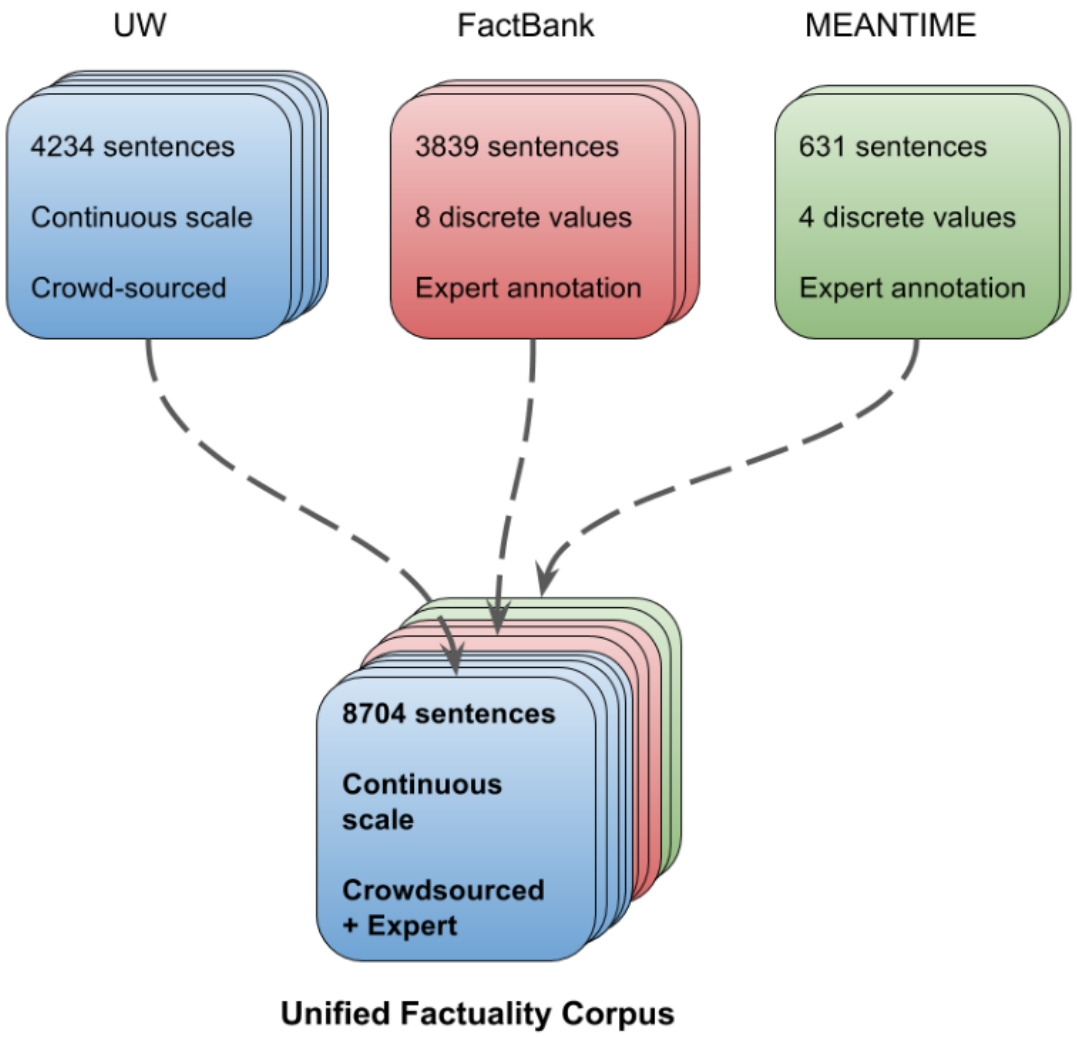
Factuality Prediction

- Previous models developed for specific datasets

→ Non-comparable results

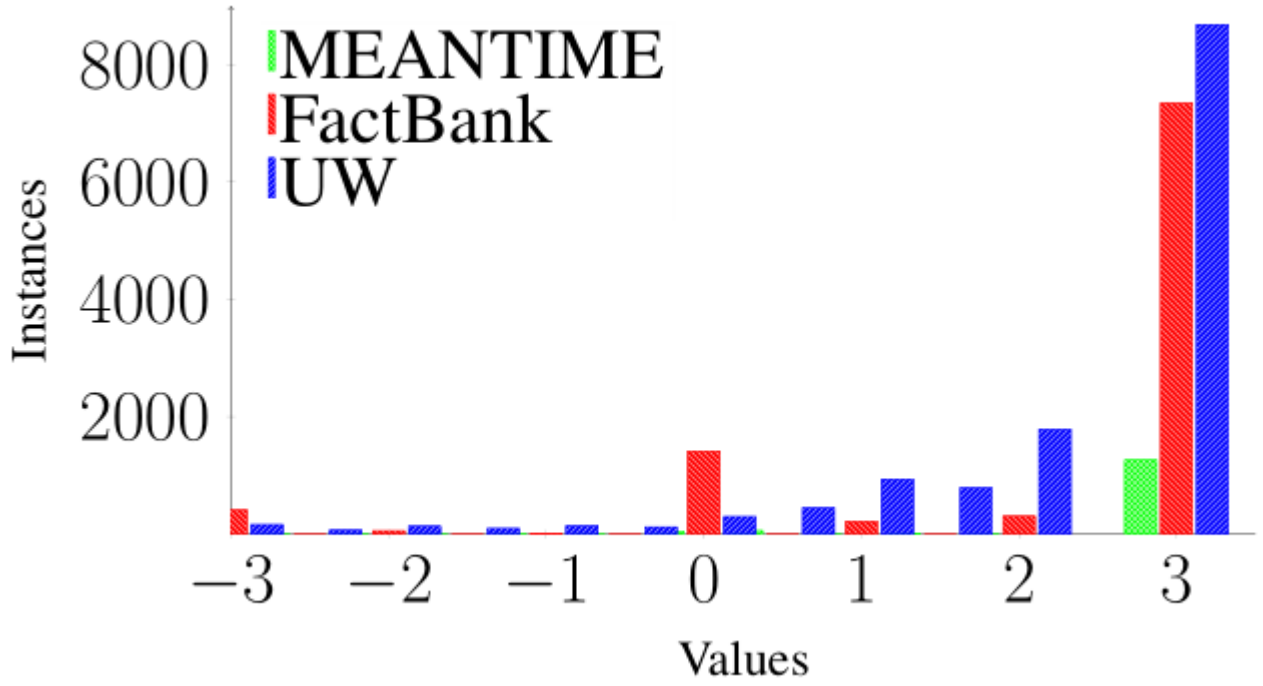
→ Limited portability

Normalizing Annotations



Biased Distribution

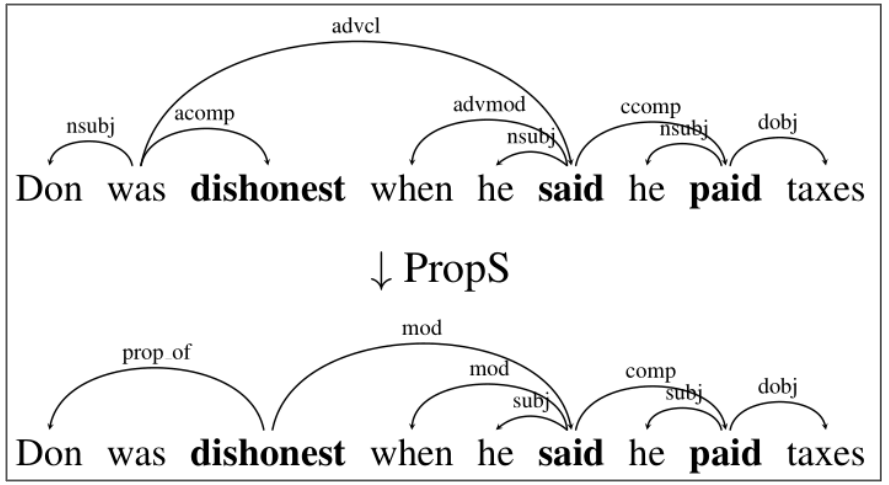
- Corpus skewed towards factual
- Inherent trait of the news domain?



Predicting

- TruthTeller (Lotan et al., 2013)
  - Used a lexicon based approach on dependency trees
  - Applied Karttunen implicative signatures to calculate factuality

- Extensions
  - Semi automatic extension of lexicon by 40%
  - Application of implicative signatures on PropS
  - Supervised learning



Evaluations

Dataset	**FactBank**		**UW**		**MEANTIME**	
	MAE	r	MAE	r	MAE	r
All-factual	.80	0	.78	0	**.31**	0
UW feat.	.81	.66	.51	**.71**	.56	.33
AMR	.66	.66	.64	.58	.44	.30
Rule-based	.75	.62	.72	.63	.35	.23
Supervised	**.59**	**.71**	**.42**	.66	.34	**.47**

Evaluations

Dataset	FactBank		UW		MEANTIME	
	MAE	r	MAE	r	MAE	r
All-factual	.80	0	.78	0	**.31**	0
UW feat.	.81	.66	.51	**.71**	.56	.33
AMR	.66	.66	.64	.58	.44	.30
Rule-based	.75	.62	.72	.63	.35	.23
Supervised	**.59**	**.71**	**.42**	.66	.34	**.47**

Marking all propositions as factual
Is a strong baseline on this dataset

Evaluations

Dataset	FactBank		UW		MEANTIME	
	MAE	r	MAE	r	MAE	r
All-factual	.80	0	.78	0	**.31**	0
UW feat.	.81	.66	.51	**.71**	.56	.33
AMR	.66	.66	.64	.58	.44	.30
Rule-based	.75	.62	.72	.63	.35	.23
Supervised	**.59**	**.71**	**.42**	.66	.34	**.47**

Dependency features correlate well

Evaluations

Dataset	FactBank		UW		MEANTIME	
	MAE	r	MAE	r	MAE	r
All-factual	.80	0	.78	0	**.31**	0
UW feat.	.81	.66	.51	**.71**	.56	.33
AMR	.66	.66	.64	.58	.44	.30
Rule-based	.75	.62	.72	.63	.35	.23
Supervised	**.59**	**.71**	**.42**	.66	.34	**.47**

Applying implicative signatures on AMR did not work well

Evaluations

Dataset	FactBank		UW		MEANTIME	
	MAE	r	MAE	r	MAE	r
All-factual	.80	0	.78	0	**.31**	0
UW feat.	.81	.66	.51	**.71**	.56	.33
AMR	.66	.66	.64	.58	.44	.30
Rule-based	.75	.62	.72	.63	.35	.23
Supervised	**.59**	**.71**	**.42**	.66	.34	**.47**

Our extension of TruthTeller gets good results across all datasets

Conclusions and Future Work

- Unified Factuality corpus made publicly available
  - Future work can annotate different domains

- External signal improves performance across datasets

- Try our online demo: http://u.cs.biu.ac.il/~stanovg/factuality.html

Acquiring Predicate Paraphrases from News Tweets

Vered Shwartz, Gabriel Stanovsky, and Ido Dagan
*SEM 2017

Bar-Ilan University

Acquiring Predicate Paraphrases from News Tweets

Vered Shwartz, Gabriel Stanovsky and Ido Dagan

*SEM 2017

Motivation

- Identifying that different predicate mentions refer to the same event

  e.g. in question answering:

  - Question
    - “When did **same-sex marriage** **become legal** in the US?"
  - Candidate Passages
    - “In June 2015, the Supreme Court **ruled for** **same-sex marriage**.”
    - “President Trump might **end** **same-sex marriage** next year.”

Our Contribution

- We released a resource of predicate paraphrases that we extracted automatically from news headlines in Twitter:
  - Up to 86% accuracy for predicate paraphrases At different support levels
  - Ever-growing resource: currently around 0.5 million predicate paraphrases
  - Expected to reach 2 million in a year

[a]$_0$ introduce [a]$_1$	[a]$_0$ welcome [a]$_1$
[a]$_0$ appoint [a]$_1$	[a]$_0$ to become [a]$_1$
[a]$_0$ die at [a]$_1$	[a]$_0$ pass away at [a]$_1$
[a]$_0$ hit [a]$_1$	[a]$_0$ sink to [a]$_1$
[a]$_0$ be investigate [a]$_1$	[a]$_0$ be probe [a]$_1$
[a]$_0$ eliminate [a]$_1$	[a]$_0$ slash [a]$_1$
[a]$_0$ announce [a]$_1$	[a]$_0$ unveil [a]$_1$
[a]$_0$ quit after [a]$_1$	[a]$_0$ resign after [a]$_1$
[a]$_0$ announce as [a]$_1$	[a]$_0$ to become [a]$_1$
[a]$_0$ threaten [a]$_1$	[a]$_0$ warn [a]$_1$
[a]$_0$ die at [a]$_1$	[a]$_0$ live until [a]$_1$
[a]$_0$ double down on [a]$_1$	[a]$_0$ stand by [a]$_1$
[a]$_0$ kill [a]$_1$	[a]$_0$ shoot [a]$_1$
[a]$_0$ approve [a]$_1$	[a]$_0$ pass [a]$_1$
[a]$_0$ would be cut under [a]$_1$	[a]$_1$ slash [a]$_0$
seize [a]$_0$ at [a]$_1$	to grab [a]$_0$ at [a]$_1$

https://github.com/vered1986/Chirps

Outline

- Resource creation
  - Obtaining News Tweets
  - Proposition Extraction
  - Generating Paraphrase Instances
  - Generating Paraphrase Types
- Analysis
  - Accuracy by score
  - Accuracy by time
- Comparison to existing resources

Method

Presumptions

- **Main assumption**: redundant news headlines of the same event are likely to describe it with different words.
  - This idea has been leveraged in previous work (e.g. Shinyama et al., 2002; Barzilay and Lee, 2003).

- Other assumption (this work): propositions extracted from tweets **discussing news events**, published **on the same day**, that **agree on the arguments**, are *predicate paraphrases*.
  - Let's look at some examples.

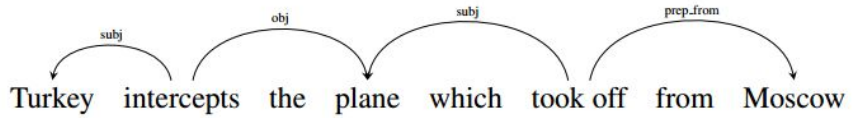
[Amazon] **to buy / is buying / to acquire** [Whole Foods]

Step #1 - Collecting News Tweets

- We query the Twitter Search API
- We use Twitter's news filter
  - Retrieves tweets containing links to news websites
- We limit the search to English tweets
- We "clean" the tweets, e.g.:
  - Remove "RT"
  - Remove links
  - Remove mentions

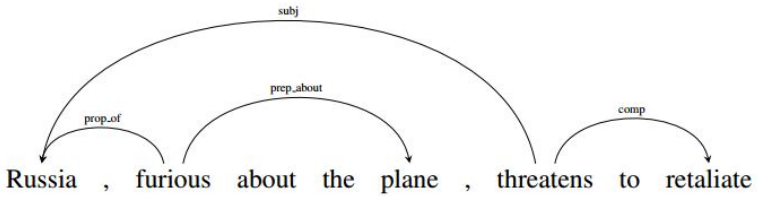
Step #2 - Proposition Extraction

- We extract propositions from the tweets using PropS (Stanovsky et al., 2016).
- We focus on binary verbal predicates, and obtain predicate templates, e.g.:



(1) [Turkey]$_0$ intercepts [plane]$_1$ (2) [plane]$_0$ took off from [Moscow]$_1$

- We employ a pre-trained argument reduction model to remove non-restrictive argument modifications (Stanovsky and Dagan, 2016).



[Russia]$_0$ threatens to [retaliate]$_1$

Step #3 - Generating Paraphrase Instances

- We consider two predicates as paraphrases if:
  1. They appear on the same day
  2. Each of their arguments aligns with a unique argument in the other predicate
- Two levels of argument matching:
  - *Strict*: short edit distance, abbreviations, etc.
  - *Loose*: partial token matching or WordNet synonyms

- Example:

Manafort hid payments from Ukraine party with Moscow ties	**[a]$_0$ hide [a]$_1$**	Paul Manafort	payments
Manafort laundered the payments through Belize	**[a]$_0$ launder [a]$_1$**	Manafort	payments
Send immigration judges to cities to speed up deportations	**to send [a]$_0$ to [a]$_1$**	immigration judges	cities
Immigration judges headed to 12 cities to speed up deportations	**[a]$_0$ headed to [a]$_1$**	immigration judges	12 cities

Step #4 - Generating Types

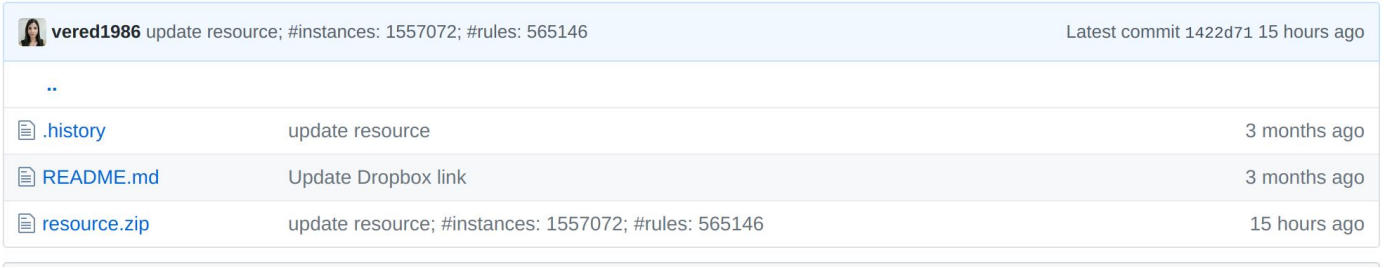
- We assign a heuristic score for each predicate paraphrase type:

$$s = count \cdot \left(1 + \frac{d}{N}\right)$$

- For example:
  - P1 = $[a]_0$ purchase $[a]_1$, P2 = $[a]_0$ acquire $[a]_1$
  - Appeared with (Amazon, Whole Foods), (Intel, Mobileye), etc. **count** times in **d** days
  - Days since resource collection begun: **N**
- **count** assigns high scores for frequent paraphrases
- **d/N** eliminates noise from two arguments participating in different events on the same day
  - e.g. 1) *Last year when Chuck Berry turned 90;* 2) *Chuck Berry dies at 90*

Resource Release

- We release our resource daily: https://github.com/vered1986/Chirps/tree/master/resource



vered1986 update resource; #instances: 1557072; #rules: 565146	Latest commit 1422d71 15 hours ago	
..		
.history	update resource	3 months ago
README.md	Update Dropbox link	3 months ago
resource.zip	update resource; #instances: 1557072; #rules: 565146	15 hours ago

- The resource release consists of two files:
  - **Instances**: predicates, arguments and tweet IDs
  - **Types**: predicate paraphrase pair types ranked in a descending order according to a heuristic accuracy score

Analysis

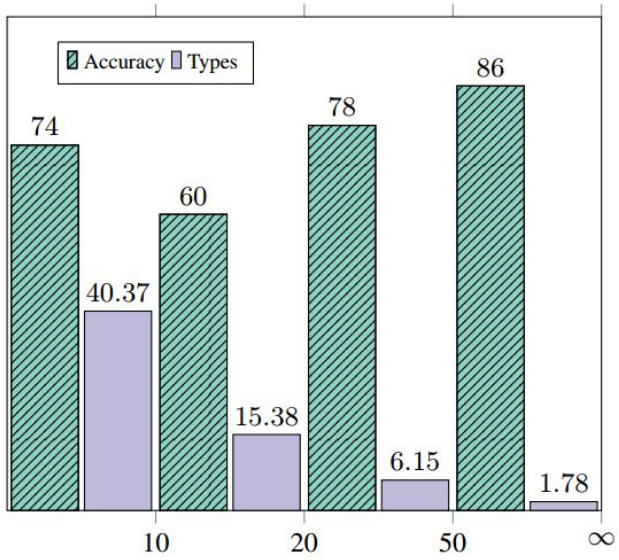
Measuring Accuracy

- We annotate a sample of the extractions using Mechanical Turk
- We follow the instance-based evaluation (Szpektor et al., 2007)
  - Judge the correctness of a paraphrase through 5 instances
  - Paraphrases are difficult to judge out-of-context

Accuracy by Score

- We partition the types into four score bins
  - Only paraphrases with at least 5 instances
- We annotate 50 types from each bin

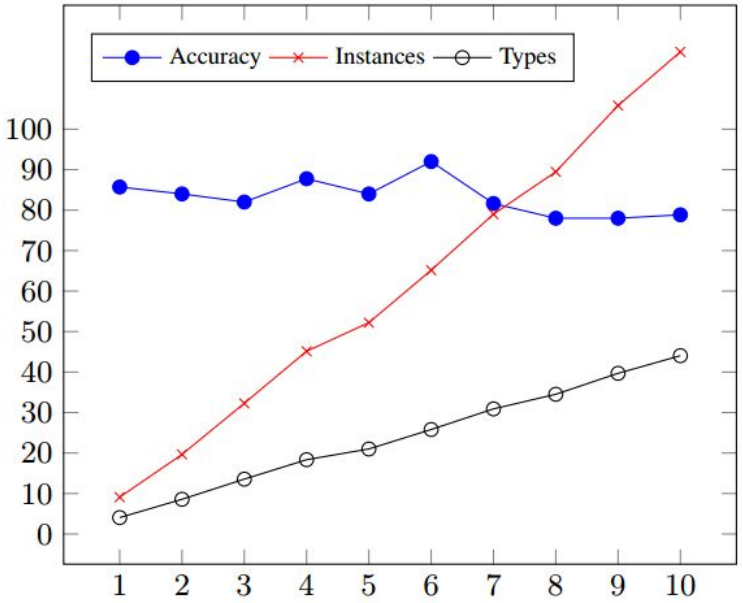
- Best scoring bin achieves up to 86% accuracy
- Accuracy generally increases with score
- Lowest-score bin contains rare paraphrases



(a) Estimated accuracy (%) and number of types ($\times 1K$) of predicate pairs *with at least 5 instances* in different score bins.

Accuracy by Time

- We estimated accuracy through each week
  - In the first 10 weeks of collection
- Accuracy at a specific time:
  - Annotating a sample of 50 predicate pair types
  - with accuracy score ≥ 20
  - in the resource obtained at that time
- Resource maintains around 80% accuracy
- We predict that the resource will contain around 2 million types in one year.



(b) Estimated accuracy (%), number of instances ($\times 10K$) and types ($\times 10K$) in the first 10 weeks.

Comparison to Existing Resources

Existing Resources

- We compare our resource with two relevant resources:
  - The Paraphrase Database (PPDB) (Ganitkevitch et al., 2013; Pavlick et al., 2015)
    - a huge collection of paraphrases extracted from bilingual parallel corpora
    - syntactic paraphrases include predicates with non-terminals as arguments
  - Berant (2012):
    - 52 million directional entailment rules
    - e.g. $[a]_0$ shoot $[a]_1 \rightarrow [a]_0$ kill $[a]_1$

Comparison to Existing Resources

- At this stage, our resource is much smaller than existing resource
  - It is infeasible to evaluate it on an evaluation set

- Our resources adds value to the existing resources:
  - 67% of the accurate types (score ≥ 50) are not in Berant
  - 62% not in PPDB
  - 49% not in neither (see table)

- Our resource contains:
  - Non-consecutive predicates
    e.g. reveal $[a]_0$ to $[a]_1$ / share $[a]_0$ with $[a]_1$

  - Context-specific paraphrases:
    e.g. $[a]_0$ get $[a]_1$ / $[a]_0$ sentence to $[a]_1$

drag $[a]_0$ from $[a]_1$	$[a]_0$ remove from $[a]_1$
leak $[a]_0$ to $[a]_1$	to share $[a]_0$ with $[a]_1$
oust $[a]_0$ from $[a]_1$	$[a]_0$ be force out at $[a]_1$
reveal $[a]_0$ to $[a]_1$	share $[a]_0$ with $[a]_1$
$[a]_0$ add $[a]_1$	$[a]_0$ beef up $[a]_1$
$[a]_0$ admit to $[a]_1$	$[a]_0$ will attend $[a]_1$
$[a]_0$ announce as $[a]_1$	$[a]_0$ to become $[a]_1$
$[a]_0$ arrest in $[a]_1$	$[a]_0$ charge in $[a]_1$
$[a]_0$ attack $[a]_1$	$[a]_0$ clash with $[a]_1$
$[a]_0$ be force out at $[a]_1$	$[a]_0$ have be fire from $[a]_1$
$[a]_0$ eliminate $[a]_1$	$[a]_0$ slash $[a]_1$
$[a]_0$ face $[a]_1$	$[a]_0$ hit with $[a]_1$
$[a]_0$ mock $[a]_1$	$[a]_0$ troll $[a]_1$
$[a]_0$ open up about $[a]_1$	$[a]_0$ reveal $[a]_1$
$[a]_0$ get $[a]_1$	$[a]_0$ sentence to $[a]_1$

Thank you!

References

[1] Yusuke Shinyama, Satoshi Sekine, and Kiyoshi Sudo. 2002. Automatic paraphrase acquisition from news articles. In Proceedings of the second international conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc., pages 313–318.

[2] Regina Barzilay and Lillian Lee. 2003. Learning to paraphrase: An unsupervised approach using multiple-sequence alignment. In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. http://aclweb.org/anthology/N03-1003.

[3] Gabriel Stanovsky, Jessica Ficler, Ido Dagan, and Yoav Goldberg. 2016. Getting more out of syntax with props. CoRR abs/1603.01648. http://arxiv.org/abs/1603.01648.

[4] Gabriel Stanovsky and Ido Dagan. 2016. Annotating and predicting non-restrictive noun phrase modifications. In Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics (ACL 2016).

[5] Idan Szpektor, Eyal Shnarch, and Ido Dagan. 2007. Instance-based evaluation of entailment rule acquisition. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. Association for Computational Linguistics, pages 456– 463. http://aclweb.org/anthology/P07-1058.

[6] Jonathan Berant. 2012. Global Learning of Textual Entailment Graphs. Ph.D. thesis, Tel Aviv University.

[7] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, pages 758–764. http://aclweb.org/anthology/N13-1092.

[8] Ellie Pavlick, Pushpendre Rastogi, Juri Ganitkevitch, Benjamin Van Durme, and Chris CallisonBurch. 2015. PPDB 2.0: Better paraphrase ranking, fine-grained entailment relations, word embeddings, and style classification. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Association for Computational Linguistics, pages 425–430. https://doi.org/10.3115/v1/P15-2070.