Event Detection and Factuality Assessment with Non-Expert Supervision

Kenton Lee, Yoav Artzi, Yejin Choi, and Luke Zettlemoyer
University of Washington

What Happened?

Nashua Corp., rumored a potential takeover target for six months, said that a Dutch company has sought U.S. approval to buy up to 25% of Nashua's shares.

What Happened?

Nashua Corp., rumored a potential takeover target for six months, said that a Dutch company has sought U.S. approval to buy up to 25% of Nashua's shares.

Event Head	Argument #1	Argument #2
rumor	-	takeover
takeover	-	Corp.
said	Corp.	sought
sought	company	approval
approval	U.S.	buy
buy	company	shares

Event Factuality

Nashua Corp., rumored a potential takeover target for six months, said that a Dutch company has sought U.S. approval to buy up to 25% of Nashua's shares.

Event Head	Argument #1	Argument #2	Factuality
rumor	-	takeover	happened
takeover	-	Corp.	did not happen
said	Corp.	sought	happened
sought	company	approval	happened
approval	U.S.	buy	did not happen
buy	company	shares	did not happen

Scalar Event Factuality

Nashua Corp., rumored a potential takeover target for six months, said that a Dutch company has sought U.S. approval to buy up to 25% of Nashua's shares.

Event Head	Argument #1	Argument #2	Factuality
rumor	-	takeover	3.0
takeover	-	Corp.	1.0
said	Corp.	sought	3.0
sought	company	approval	2.1
approval	U.S.	buy	1.5
buy	company	shares	1.2

Data Annotation

- Annotation:
  - Label the head of each event.
  - Label the factuality of event mention from the author's point of view.
- Goals:
  - Scalable to non-experts.
  - Minimal jargon in instructions.
  - Example driven.

Annotating Events

Instructions

- We are interested in finding **events** that are mentioned in sentences.
- We consider **events** to be things that may or may not occur either in the past, present or future (e.g. *earthquake, meeting, jumping, talking*, etc.)
- In some cases, it is not so clear whether a word is referring to an event or not. Consider these harder cases to be events.

Meantime , FBI agents and Metropolitan Police officers assigned to a joint terrorism task force here scanned the crowd of anti-abortion protesters at the annual March for Life on Capitol Hill , because Kopp has been either a participant in or arrested at this march in each of the last three years , according to another law enforcement official , Both officials requested anonymity .

Does the highlighted word ('scanned') refer to an event?

- ○ **Yes**
- ○ **No**

Annotating Factuality

Instructions

- Read the sentence in the box carefully.
- Rate on a scale from 3 to -3, how likely the highlighted event did or will happen **according to the author of the sentence**
- Since this task is somewhat subjective, we will try to be lenient when scoring test questions.
- If this is your first time doing this task, please read the following examples and explanations.

While Woods was pleased with his results , he was n't proud of the way he had to scramble on one of the easier courses on the PGA Tour .

On a scale from 3 to -3, how certain is the author that the highlighted event: 'scramble' did or will happen?

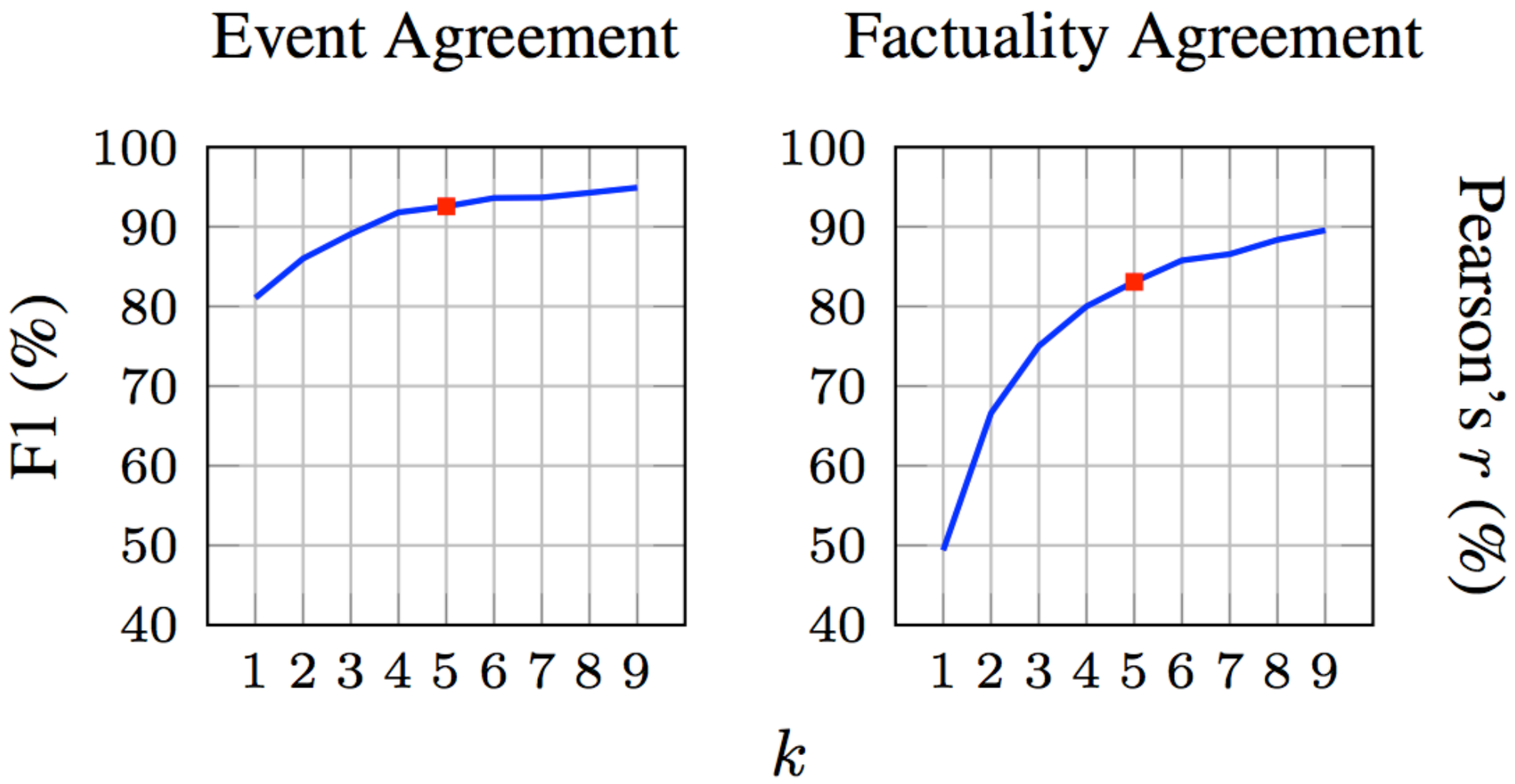
- 3 : The author believes that it **certainly** did or will happen.
- 2
- 1
- 0 : The author's stance is **neutral**.
- -1
- -2
- -3 : The author believes that it **certainly** did **not** or will **not** happen.

Example Annotations

U.S. embassies and military installations around the world were ordered[3.0] to set[2.6] up barriers and tighten[2.6] security systems to prevent[1.8] easy access[-2.4] by unauthorized people --Americans and foreigners.

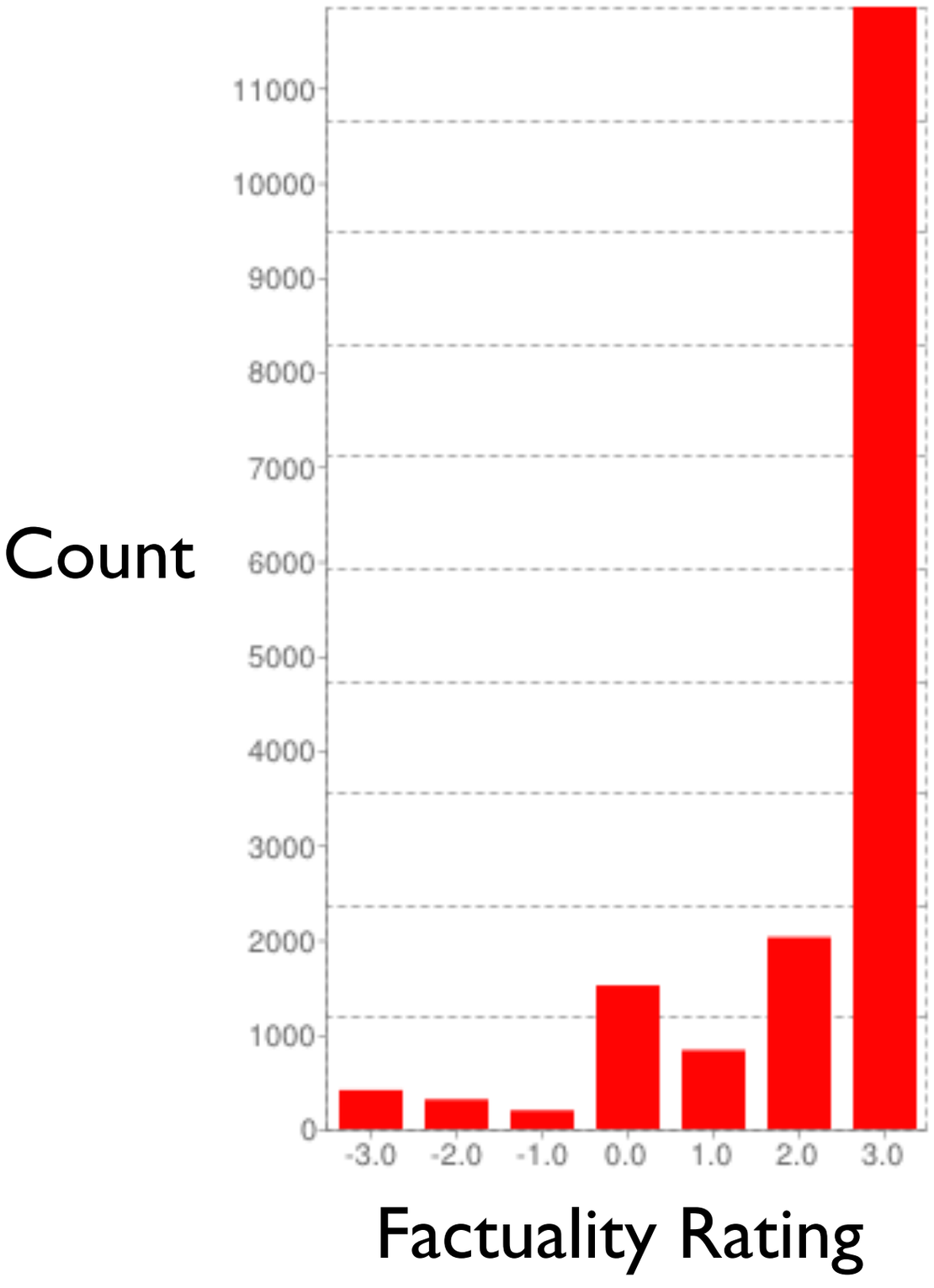
The White House said[3.0] President Bush has approved[3.0] duty-free treatment[1.6] for imports[2.8] of certain types of watches that aren't produced[0.0] in “significant quantities” in the U.S., the Virgin Islands and other U.S. possessions.

Meta-annotator Agreement



Pairwise agreement statistics vs. the number of judgments for each meta-annotator.

Factuality Bias in Newswire



Histogram of factuality ratings from the TempEval-3 corpus.

Comparison to FactBank

Discretized ratings	FactBank Labels								
	CT-	PR-	PS-	PS+	PR+	CT+	CTu	NA	Uu
-3	39	0	0	0	0	0	0	0	29
-2	29	2	0	0	0	0	0	0	44
-1	16	4	1	0	0	3	0	0	58
0	15	0	5	2	0	7	0	1	95
1	7	0	1	30	4	27	2	0	337
2	4	1	0	20	42	260	0	0	564
3	2	0	0	1	10	2760	0	0	771

Confusion matrix between our discretized labels and factuality categories from FactBank (Sauri and Pustejovsky, 2009)

Modeling Factuality

Objective :

Hybrid of Support Vector Regression and the LASSO

$$\min_{w} \|w\|_1 + C \sum_{i=1}^{N} \max(0, |y_i - w^\top \phi(x_i)| - \epsilon)$$

Features :

- Lemma of the target event.
- Part-of-speech of the target event.
- Dependency paths of up to length 2 from the target event.

Dependency Representation

John did not expect to **return**.

Capture event-event interactions through dependency paths:

$$\textit{not} \leftarrow [\textbf{neg}] — \textit{expect} — [\textbf{xcomp}] \rightarrow \langle * \rangle$$
$$\langle * \rangle \leftarrow [\textbf{neg}] — \textit{expect} — [\textbf{xcomp}] \rightarrow \langle * \rangle$$
$$\textit{not} \leftarrow [\textbf{neg}] — \langle * \rangle — [\textbf{xcomp}] \rightarrow \langle * \rangle$$
$$\langle * \rangle \leftarrow [\textbf{neg}] — \langle * \rangle — [\textbf{xcomp}] \rightarrow \langle * \rangle$$

Results

Model	Dev.			Test		
	P	R	F1	P	R	F1
Our system	**90.1**	**90.9**	**90.5**	85.5*	**87.8**	**86.6**
NAVYTIME	84.7*	79.6*	82.1*	**87.7**	78.3*	82.7*

Figure 5: Results for the detection task.

Model	Dev.		Test	
	MAE	r	MAE	r
Our system	**46.2**	**74.9**	**51.1**	**70.8**
SVR	50.3*	74.8	57.1*	69.4
DISCRETE	50.3*	68.6*	52.4	62.2*
DIAB	58.7*	51.1*	62.0*	50.8*

Figure 6: Results for the factuality task.

Common Errors

Missing lexical cues (64%)

Wong Kwan will be lucky to **break** even.

Long-distance inference (16%)

Mesa had rejected a general proposal from StatesWest to **combine** the two carriers.

World knowledge and pragmatics (12%)

There was no hint of **trouble** in the last conversation between controllers and TWA pilot Steven Snyder.

Future Work

- Active learning for efficient lexical coverage.
- Joint models to better capture event-event interactions.
- Extrinsic evaluation with information extraction.