A TACRED Data Collection and Validation

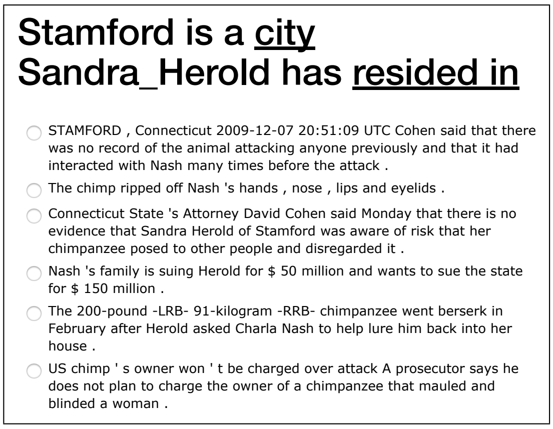
In this appendix, we describe the way we collect and validate TACRED in full detail.

A.1 Data Collection

TACRED leverages the work done selecting query entities and annotating system responses in the TAC KBP evaluations. In each year of the TAC KBP evaluation (2009–2015), 100 query entities are given to participating KBP systems with the aim of filling in valid knowledge base entries for these entities. Our annotation effort re-uses these query entities, annotating each sentence in the source corpus that contains one of these entities. Given the set of mention pairs (e.g., *Penner* and *Lisa Dillman*) containing an evaluation entity, the mention pair can have either 1) been extracted during a previous KBP competition and marked correct by an LDC annotator, or 2) been generated automatically from candidate mention pairs in the corpus. For clarity, we refer to the former as **LDC examples** and the latter as **generated examples**, and describe them separately.

LDC examples. For examples in this category, although the relations have been annotated by an LDC annotator, the provenance for the mention pairs provided in TAC KBP evaluation files are often too general or imprecise; for example in early years only the document that contains a mention pair is given as provenance. We solve this problem with a two-stage annotation task (HIT) in Mechanical Turk: In the first task, Turk annotators are provided with the mention pair and its relation (annotated by LDC), and asked to find a sentence in the document that expresses the extraction. In the second task, annotators are asked to identify the spans of both the subject and object entities. See Figure 7 and Figure 8 for example interfaces provided to Turk annotators.

Generated examples. To further collect examples that are not annotated by LDC, we first run annotations on the corpus using a combination of Stanford's statistical coreference system (Clark and Manning, 2015) and the Illinois Wikifier (Ratinov et al., 2011). Then we collect all mention pairs in which one mention is linked to one of the query entities by the entity linker. To prevent the resulting dataset from being skewed towards commonly occurring query entities such as “Barack Obama”, we enforce a hard upper limit on the number of collected mention pairs containing a query entity. Specifically, for each query entity q, we retrieve N_q sentences from the KBP corpus that contain an entity mention linked to q. Then let N_{q^c} denote the number of extractions submitted by competing KBP systems that were also deemed correct by human annotators, we want N_q to be proportional to N_{q^c}, and heuristically set: $N_q = \min(9 \cdot N_{q^c}, 300)$. Next, each mention pair, along with the corresponding sentence in which it occurs, is annotated for its relation type (or *no_relation*) as a task on Mechanical Turk. Figure 9 shows an example task interface for generated examples on Mechanical Turk.



Stamford is a city
Sandra_Herold has resided in

○ STAMFORD , Connecticut 2009-12-07 20:51:09 UTC Cohen said that there was no record of the animal attacking anyone previously and that it had interacted with Nash many times before the attack .

○ The chimp ripped off Nash 's hands , nose , lips and eyelids .

○ Connecticut State 's Attorney David Cohen said Monday that there is no evidence that Sandra Herold of Stamford was aware of risk that her chimpanzee posed to other people and disregarded it .

○ Nash 's family is suing Herold for $ 50 million and wants to sue the state for $ 150 million .

○ The 200-pound -LRB- 91-kilogram -RRB- chimpanzee went berserk in February after Herold asked Charla Nash to help lure him back into her house .

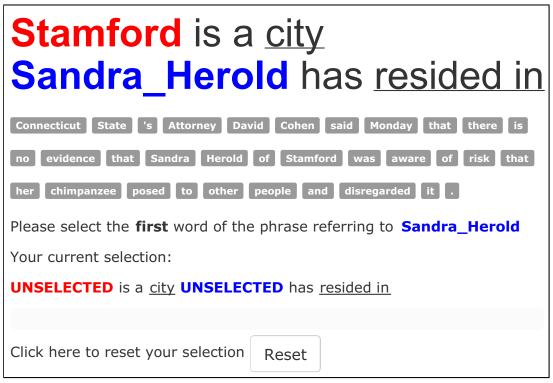
○ US chimp ' s owner won ' t be charged over attack A prosecutor says he does not plan to charge the owner of a chimpanzee that mauled and blinded a woman .

Figure 7: Example of an LDC examples HIT on Mechanical Turk for identifying the relevant sentence. The annotator is presented with every sentence from the document as well as the extraction for which to find the sentence.

A.2 Data Validation

In order to maintain the quality of TACRED, we validate the collected data both during and after the annotation process. We made use of crowdsourced data from a previous annotation effort on the same relation set (Angeli et al., 2014a). During annotation, 10% of the HITs presented to a worker are sanity check examples from this previous data, and annotators whose error rate on these examples exceeds 25% were asked to have their work reannotated.

After the data collection is done, one of the authors manually examined 300 sampled instances. The estimated annotation accuracy is 93.3%, with a confidence interval of (89.9%, 95.9%). In addition, for the collected generated examples, we estimate inter-annotator agreement using 761 sampled



Stamford is a city
Sandra_Herold has resided in

Connecticut State 's Attorney David Cohen said Monday that there is no evidence that Sandra Herold of Stamford was aware of risk that her chimpanzee posed to other people and disregarded it .

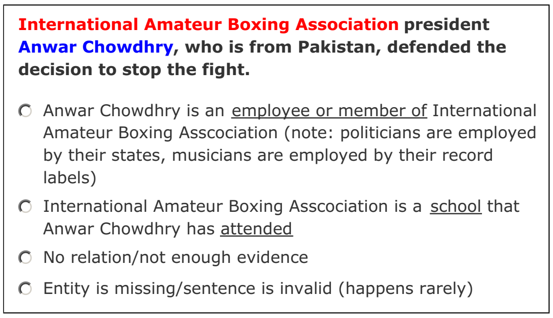
Please select the **first** word of the phrase referring to **Sandra_Herold**

Your current selection:

UNSELECTED is a city **UNSELECTED** has resided in

Click here to reset your selection

Figure 8: Example of an LDC examples HIT on Mechanical Turk for identifying the mention spans. The annotator is presented with a sentence obtained from the HIT shown in Figure 7 as well as the corresponding extraction and asked to identify the spans of the subject and object mentions in the extraction.



International Amateur Boxing Association president Anwar Chowdhry, who is from Pakistan, defended the decision to stop the fight.

○ Anwar Chowdhry is an employee or member of International Amateur Boxing Association (note: politicians are employed by their states, musicians are employed by their record labels)

○ International Amateur Boxing Asscociation is a school that Anwar Chowdhry has attended

○ No relation/not enough evidence

○ Entity is missing/sentence is invalid (happens rarely)

Figure 9: Example of a generated examples HIT. The subject entity is highlighted in blue and the object entity is highlighted in red. The annotator is asked to select among a set of plausible relations that are compatible with the subject and object entity types, along with an option to state that none of the presented relations hold.

mention pairs shown to five annotators. Results are shown in Table 7.

Metric	Score
5 annotators agree	74.2%
$\geq$ 4 annotators agree	90.5%
$\geq$ 3 annotators agree	100.0%
Fleiss Kappa	54.4%

Table 7: Estimated inter-annotator agreement using 761 sampled mention pairs.

A.3 Data Statistics

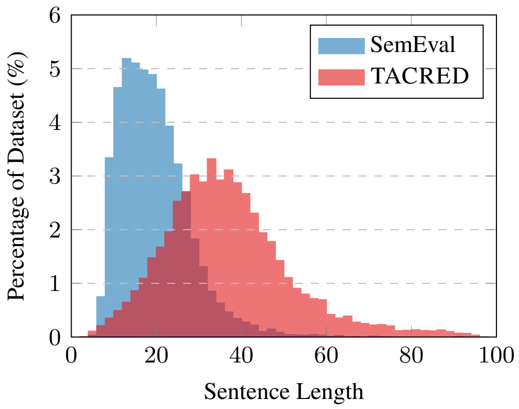


Figure 10: Distribution of sentence lengths in SemEval 2010 task 8 and TACRED.

In total, we collect 10,691 annotations from the LDC examples task and 110,021 annotations from the generated examples task. After removing examples where the subject and object entities overlap, we arrive at a total of 119,474 examples. About 78.7% of all examples are annotated as *no_relation*, which we showed to be crucial for training high-precision relation extraction models for the TAC KBP 2015 slot filling evaluation. Furthermore, we find that sentences in TACRED tend to be much longer than in the SemEval dataset (Figure 10).

Table 8 presents detailed statistics on this dataset. We also include sampled training examples in Table 9.

B Model Training Details

Here we describe the way we train our models in detail for replicability.

Model hyperparameters. We use 200 for word embedding size and 30 for every other embedding (i.e., position, POS or NER) size. For CNN models, we use filter window sizes ranging from 2 to 5, and 500 filters for each window size. For the SDP-LSTM model, in addition to POS and NER embeddings, we also include the type of dependency edges as an additional embedding channel. For our proposed position-aware neural sequence model, we use attention size of 200. For all models that require LSTM layers, we find a 2-layer stacked LSTMs works better than a single-layer LSTM. We use one-directional LSTM layers in all of our experiments. Empirically we find bi-directional LSTM layers give no improvement to our proposed position-aware sequence model and marginal improvement to the simple LSTM model. We do not add max-pooling layers after

LSTM layers as we find this harms the performance.

Training. During training, we employ standard dropout (Srivastava et al., 2014) for CNN models, and RNN dropout (Zaremba et al., 2014) for LSTM models. Additionally, for CNN models we apply ℓ_2 regularization with coefficient 10^{-3} to all filters to avoid overfitting. We use *AdaGrad* (Duchi et al., 2011) with a learning rate of 0.1 for CNN models and 1.0 for all other models. We train CNN models for 50 epochs and other models for 30 epochs, with a mini-batch size of 50. We monitor the training process by looking at the micro-averaged F_1 score on the dev set. Starting from the 20th epoch, we decrease the learning rate with a decay rate of 0.9 if the dev set micro-averaged F_1 score does not increase after every epoch. Finally, we evaluate the model that achieves the best dev set F_1 score on the test set.

Relation	Total	Percentage	Train	Development	Test
			2009–2012	2013	2014
no_relation	94001	78.68%	60179	19305	14517
org:alternate_names	1515	1.27%	893	380	242
org:city_of_headquarters	656	0.55%	437	125	94
org:country_of_headquarters	878	0.73%	540	215	123
org:dissolved	41	0.03%	29	8	4
org:founded	199	0.17%	103	49	47
org:founded_by	343	0.29%	145	109	89
org:member_of	222	0.19%	147	39	36
org:members	330	0.28%	194	95	41
org:number_of_employees/members	144	0.12%	87	35	22
org:parents	528	0.44%	332	120	76
org:political/religious_affiliation	148	0.12%	118	13	17
org:shareholders	168	0.14%	87	66	15
org:stateorprovince_of_headquarters	407	0.34%	266	83	58
org:subsidiaries	516	0.43%	326	138	52
org:top_members/employees	3182	2.66%	2138	635	409
org:website	302	0.25%	133	133	36
per:age	977	0.82%	416	292	269
per:alternate_names	172	0.14%	111	48	13
per:cause_of_death	384	0.32%	127	199	58
per:charges	322	0.27%	77	120	125
per:children	385	0.32%	235	109	41
per:cities_of_residence	857	0.72%	421	203	233
per:city_of_birth	126	0.11%	77	40	9
per:city_of_death	271	0.23%	102	133	36
per:countries_of_residence	978	0.82%	498	281	199
per:country_of_birth	74	0.06%	39	26	9
per:country_of_death	83	0.07%	10	57	16
per:date_of_birth	127	0.11%	78	39	10
per:date_of_death	451	0.38%	151	238	62
per:employee_of	2621	2.19%	1837	433	351
per:origin	794	0.66%	373	257	164
per:other_family	417	0.35%	233	96	88
per:parents	334	0.28%	164	59	111
per:religion	186	0.16%	61	65	60
per:schools_attended	277	0.23%	178	62	37
per:siblings	284	0.24%	178	37	69
per:spouse	569	0.48%	311	185	73
per:stateorprovince_of_birth	88	0.07%	47	30	11
per:stateorprovince_of_death	133	0.11%	65	53	15
per:stateorprovinces_of_residence	560	0.47%	374	89	97
per:title	4424	3.70%	2733	1065	626
Total	119474	100.00%	75050	25764	18660

Table 8: Relation distribution of the TACRED dataset.

Example Sentences	Subject Type	Object Type	Relation Labels
Carey will succeed **Cathleen P. Black**, who held the position for 15 years and will take on a new role as *chairwoman* of Hearst Magazines, the company said.	Person	Title	per:title
Baldwin declined further comment, and said JetBlue chief *executive* Dave Barger was unavailable.	Person	Title	no_relation
Irene Morgan Kirkaldy, who was born and reared in Baltimore, lived on Long Island and ran a child-care center in Queens with her second husband, *Stanley Kirkaldy*.	Person	Person	per:spouse
Cummings, current holder of the Seventh District seat held by *Mr. Mitchell*, sponsored legislation last year that named a Baltimore post office in the veteran congressman's honor.	Person	Person	no_relation
Blackburn Rovers announced Tuesday they had sacked *Paul Ince* as their manager, a statement on the Premier League club's website said.	Organization	Person	org:top_members/employees
Kerry wrote a letter to *Pickens*, saying he would donate any proceeds to the **Paralyzed Veterans of America**, the Associated Press reported.	Organization	Person	no_relation
Forsberg launched the anti-nuclear movement with a paper she wrote while obtaining a doctorate in international studies at *Massachusetts Institute of Technology*.	Person	Organization	per:schools_attended
He received an undergraduate degree from Morgan State University in 1950 and applied for admission to graduate school at the *University of Maryland in College Park*.	Person	Organization	no_relation

Table 9: Sampled training examples from the TACRED dataset, with subject entity highlighted in bold and blue and object entities highlighted in italics and red.