Rationally Reappraising ATIS-based Dialogue Systems

Jingcheng Niu, Gerald Penn *

University of Toronto

Findings

- Recent neural architectures overfit when evaluated on ATIS, overshadowing any potential gain from better contextual inference.
- However, ATIS may still play a very prominent role for:
  - the development of a syntactically annotated slot-filling corpus,
  - the transfer of learning between parsers on different domains, and
  - the appropriation of such a portable parser to slot filling.
- We discovered and taxonomized a large number of errors in ATIS, and have provided a repaired version of the ATIS corpus that creates a 19 ∼ 52% relative error reduction.

Errors in ATIS

Split	Train		Test	
	total	%	total	%
total utterances	4978	100	893	100
incorrect	132	2.61	46	5.15
UNK	46	0.92	46	5.15
total slots	16561	100	2837	100
incorrect	188	1.14	65	2.29

Annotation Mistakes by Dataset

Error Taxonomy

- **Incorrect IOB Segmentation**

  B-state_name I-state_name

  List airports in Arizona Nevada and California please

- **Wrong Word Selection**

  B-city_name I-city_name I-city_name

  … the city of Boston

- **Missing Labels**

  O

  All am flights departing …

- **Concept Mistakes**

  B-fromloc.city_name B-fromloc.city_name

  All flights before 10 am Boston Denver

- **NE Mistakes**

  B-city_name B-city_name B-city_name I-city_name

  Westchester county New York

- **Out-of-Vocabulary (UNK)**

  e.g., *What is UNK?*

Split	Train		Test	
	utterances	instances	utterances	instances
IOB	2	2	2	2
W W Sel	22	22	1	1
Missing	29	30	4	4
Concept	72	120	28	46
NE	12	13	11	11
OOV	46	46	46	46

Annotation Mistakes by Type

Systems Evaluated

- **RNN:** Mesnil et al. (2013) evaluated both the Elman-type RNN and the Jordan-type RNN on ATIS. This work is mostly considered as one of the first to apply RNN on the slot-filling task.
- **LSTM:** Yao et al. (2014) built upon previous simple RNN work and investigated the effectiveness of LSTMs for the slot-filling task.
- **Encoder-Decoder:** Kurata et al. (2016) created an encoder-decoder-style system that can leverage contextual information from the whole input sequence.
- **Self-attentive BiLSTM:** Li et al. (2018) reaches the current state-of-the-art performance on ATIS by proposing a self-attentive model with a gate mechanism that utilizes sentence intent information[a].
- **Encoder-Decoder with Focus:** Zhu and Yu (2017) incorporated an attention mechanism, as well as a newly proposed focus mechanism into a bidirectional LSTM encoder-decoder system.
- **Our rule-based system:** A phrase-structure rule based system using the Attribute Logic Engine (ALE) (Carpenter and Penn, 1994). The system has an all-paths chart parser that produces phrase structure forests, a definite-clause extension that assign slot labels to tokens, and a greedy algorithm that breaks ties.

[a] For the sake of fairness to other systems, the intent information is not included in our evaluation.

Experimental Results

We evaluated the rule-based system and these 5 neural systems on 4 test sets:

- **Test**: the original ATIS test set,
- **Fixed**: our fixed ATIS test set,
- **UNK**: our fixed ATIS test set, without discarding any utterance with an UNK problem, and
- **X**: ATIS_X set from Zhu and Yu (2018) that replaces NEs in the utterance with unseen ones.

Model		Test	Fixed	UNK	X
RNN	Complete ATIS	93.56	95.83	94.71	92.3
	Full Parse	93.8	96.8	95.65	93.49
LSTM	Complete ATIS	93.86	96.47	95.54	93.29
	Full Parse	94.22	97.44	96.4	94.57
Encoder-Decoder	Complete ATIS	94.75	95.77	96.84	91.85
	Full Parse	94.89	96.49	97.55	92.74
Self-att. BiLSTM	Complete ATIS	94.87	96.99	96.05	93.60
	Full Parse	95.06	98.02	97.25	94.72
Focus	Complete ATIS	95.02	97.61	96.42	84.31
	Full Parse	95.19	98.10	96.86	83.81
Rule-Based	rand.	93.00	95.82	94.47	92.92
	scep.	90.91	94.10	92.44	90.68
	cred.	94.33	96.66	95.84	94.35
Full Parse	rand.	95.61	98.62	97.19	95.49
	scep.	94.81	97.93	96.41	94.59
	cred.	96.68	99.10	98.31	96.51
Full Parse %		80.87	81.81	80.87	80.99

NSERC CRSNG BRAINALLIANCE

*{niu,gpenn}@cs.toronto.edu