Rationally Reappraising ATIS-based Dialogue Systems

Jingcheng Niu and **Gerald Penn**
Department of Computer Science
University of Toronto
Toronto, Canada
{niu,gpenn}@cs.toronto.edu

Abstract

The Air Travel Information Service (ATIS) corpus has been the most common benchmark for evaluating Spoken Language Understanding (SLU) tasks for more than three decades since it was released. Recent state-of-the-art neural models have obtained F1-scores near 98% on the task of slot filling. We developed a rule-based grammar for the ATIS domain that achieves a 95.82% F1-score on our evaluation set. In the process, we furthermore discovered numerous shortcomings in the ATIS corpus annotation, which we have fixed.

This paper presents a detailed account of these shortcomings, our proposed repairs, our rule-based grammar and the neural slot-filling architectures associated with ATIS. We also rationally reappraise the motivations for choosing a neural architecture in view of this account. Fixing the annotation errors results in a relative error reduction of between 19.4 and 52% across all architectures. We nevertheless argue that neural models must play a different role in ATIS dialogues because of the latter's lack of variety.

1 Introduction

Slot filling has received a great deal of recent attention from the SLU community. Typically, it is characterized as a sequence labeling problem in which certain tokens are identified as fillers that contribute argument values to a meaning representation through "slot" positions in the utterance. Wang et al. (2011) first used conditional random fields (CRF) for slot filling. A few years later, inspired by the success of recurrent neural networks (RNN) in language modeling (Mikolov et al., 2011), Mesnil et al. (2013) developed the first RNN slot filler that achieved a relative error reduction of 14%. Subsequently, different variations of RNN such as LSTM (Yao et al., 2014) were developed for slot filling, followed by encoder-decoder models that could utilize information from the entire sentence (Kurata et al., 2016), both of which avail themselves of an attention mechanism (Zhu and Yu, 2017; Li et al., 2018). As recently as Wang et al. (2018), Deep Reinforcement Learning (DRL) has been proposed as a way to refine encoder-decoder models on sparsely distributed tags; this has achieved the highest reported performance so far.

This development has taken place in parallel, however, with work that has used qualitative error analyses to cast doubt on the continued use of ATIS as a benchmark for progress in slot filling. Most recently, Béchet and Raymond (2018) conclude that ATIS is simply too "shallow" to offer anything of additional substance for DNN-based architectures to achieve, formulating a three-way taxonomy of errors in the reference annotation for the ATIS corpus that account for roughly half of the remaining errors still faced by state-of-the-art slot filling models. Even prior to the recent popularity of neural architectures, Tur et al. (2010) cited a problem with earlier n-gram-based modeling approaches, which tended to fit every utterance into a known sample without regard to domain knowledge or aspects of global context that could override local n-gram contexts.

We present here: (1) a thorough taxonomy of ATIS annotation errors, reminiscent of the taxonomy of slot-filling errors in Béchet and Raymond (2018), (2) a repaired version of the ATIS reference annotation, (3) a freely available rule-based grammar of the ATIS domain,[1] that offers an alternative to a language-modeling-based approach, incorporating both domain knowledge and non-local inference as advocated for by Tur et al. (2010), (4) an experimental trial in which five recent neural architectures are evaluated on the re-

[1] http://www.ale.cs.toronto.edu/grammars/atis.pl

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 5503–5507
Florence, Italy, July 28 - August 2, 2019.

Index:105	American	airlines	leaving	Phoenix
IOB	B	I	O	B
Concept				fromloc
NE	airline_name	airline_name		city_name

Table 1: Example of an utterance in ATIS.

paired ATIS annotation alongside the rule-based grammar, and (5) an analysis of the experimental results that, while broadly supporting the conclusions of Béchet and Raymond (2018), attempts to circumscribe the possible meaning of "shallow" more precisely.

Crucial to our experimental results and our conclusions is a recent, independent modification of the ATIS corpus (Zhu and Yu, 2018) that inadvertently exposes some of what neural approaches are modeling with respect to slot fillers.

2 ATIS Corpus

2.1 Dataset

The ATIS Spoken Language Systems Pilot Corpus (Hemphill et al., 1990) contains utterances of users asking flight-related questions that could be answered by a relational query search from the ATIS database. For the task of slot filling, only the text part of the corpus is used. Generally, 4978 Class A utterances in the ATIS-2 and ATIS-3 corpora are used as the training set, and 893 utterances from ATIS-3 Nov93 and Dec94 are selected as the testing set. Developers may randomly split the 4978 utterances into a training set (for us, 90%) and a development test set (10%).

The text data are converted to the format suitable for the slot filling task. Each token of an utterance is considered to be a potential slot, and each slot should contain a tag, with an optional Concept part and a mandatory Named Entity (NE) part, in the In/Out/Begin (IOB) format. Mesnil et al. (2013) converted the relational queries into that format using an automatic process. Table 1 is an annotated example. The entire dataset contains 9 distinct concepts and 44 NEs that yield 127 total possible tags. For ease of reference, we number both the training and test sets in lexicographical order here, starting from 0.

2.2 Errors in Annotation

Béchet and Raymond (2018) identify three sources of error: annotations missing slots entirely or transposing labels, for example, between departure and arrival cities; determinately reading an utterance that is naturally ambiguous (no system should be penalized for having guessed another valid reading); and labeling only the first of several instances of the same NE in the same utterance (systems that label more than one are penalized). 1.14% of the slots in the training set are incorrectly labeled overall, as are 2.29% of those in the test set. These percentages are significant, given that state-of-the-art systems commonly report error rates of between 1.2% to 6%. Note that there are almost twice as many errors in the test set as in the training set on a percentage basis. About half of these are ambiguous slots arising from the use of "UNK" for hapax legomena. In these 46 cases, the slot cannot be determined without knowledge of what the word formerly was. Most egregiously, five of utterances 785–791 are "What is UNK?" and the other two are "What is a UNK?".

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[2]	100	2837	100
incorrect	188	1.14	65	2.29

Table 2: Annotation Mistakes by Dataset.

The test set is unique in other respects. Six of its slot labels (*B-booking_class, B-flight, B-stoploc.airport_code, I-state_name, I-flight_number* and *B-compartment*) are not found in the training set. Except for *B-stoploc.airport_code*, the other five are NE annotation errors. The test set also handles the word *noon* differently: four instances are treated as a *period_of_day*, whereas all occurrences of *noon* in the training set are treated as a *time*.

2.3 Taxonomy

We have created our own error classification (Figure 1 and Table 3). Not all of these classes map onto one of the three in Béchet and Raymond (2018). The taxonomy and errors were labelled independently by two annotators, who were then forced to reconcile where they disagreed.

3 Rule-based Grammar

In addition to repairing the ATIS annotations, we developed a rule-based grammar for use as a

[2]After fixing ATIS, there were 4932 training utterances (16419 slots) and 847 test utterances (2665) left.

- **Incorrect IOB Segmentation** In the test set, 309: *"List airports in Arizona, Nevada and California please."* unifies the two states *Arizona* and *Nevada* into one slot, and was annotated as *B-state_name* and *I-state_name*. Corrected.
- **Wrong Word Selection** Some slots select the wrong words. Utterance 1374: *"I need information on ground transportation between airport and downtown in the city of Boston"* labels the whole phrase *city of Boston* as *toloc.city_name*, whereas elsewhere only *Boston* is labeled. Chose dominant word sequence.
- **Missing Labels** Words that should be annotated are not (equivalent to label, O, i.e. outside of any slot). For example, in 29: *"All am flights departing Pittsburgh arriving Denver."*, the abbreviation *'am'* should have been labeled *B-depart_time.period_of_day*, but was not annotated. Annotation added.
- **Concept Mistakes** These are the most prevalent annotation error. For example, *"Denver"* in 40: *"All flights before 10 am Boston Denver."* was annotated as *B-fromloc.city_name*, where it should have been *toloc*. Includes ambiguities that are not consistently annotated (we chose the dominant annotation) as well as unambiguous fillers that bear more than one concept role (which the annotation standard does not permit; these were discarded).
- **NE Mistakes** These appear in both the training and the test set. For example, in utterance 29: *"Flights from Denver to Westchester county New York weekdays."*, *New York* means the state of New York, not New York City, but its NE was labeled as a *city_name* instead of *state_name*. Corrected.
- **Out-of-Vocabulary (UNK)** These are found in the training set (e.g., 4394: *"What is* $\langle unk \rangle$*?"*) and the test set, as discussed above. Discarded the utterance.

Figure 1: Taxonomical classes, examples, and repair actions taken.

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

Table 3: Annotation Mistakes by Taxonomic Class.

baseline and domain-specific knowledge source, particularly of time and location phrases. We used the Attribute Logic Engine (ALE) (Carpenter and Penn, 1994), a grammar development system and logic programming language based upon typed feature structures. ALE compiles grammars into an all-paths chart parser that produces phrase structure forests. We use the logic programming extension to project words into individual IOB slots, given a parsing chart.

The grammar does not generate a spanning parse for utterances with multiple sentences (e.g., 3612:*"US air 269 leaving Boston at 428. What is the arrival time in Baltimore?"*). These, as well as single sentences for which no spanning edge is found, are instead projected using a covering of edges that is selected with the greedy algorithm shown in Algorithm 1. This algorithm prefers longer spans to shorter spans and breaks ties by selecting one edge uniformly at random.

Algorithm 1 GREEDY(*edges*)

$long \leftarrow$ a longest edge in $edges$
$L \leftarrow$ edges finish before $long$
$R \leftarrow$ edges start after $long$
return GREEDY(L) + *long* + GREEDY(R)

The grammar uses 601 lexical entries (one or more for each of the 573 word types in ATIS), 643 feature structure types, 22 features and 330 phrase structure rules. The feature structure types that we defined were for two major purposes: 168 syntactic types that label the nodes of a parse tree, and 475 types that declare appropriate values for features. Every syntactic node label has features that refer to a list of slot fillers (TAGS) and a list of tokens (WORDS) in the subtree at which it is rooted.

Among the 330 grammar rules, 65 rules are used to capture multi-word expressions (MWE), which ALE does not otherwise support. Only 161 rules are designed specifically for ATIS, with the remaining 104 being general rules of English grammar. Nouns are further divided into different ATIS-specific slot values such as cities, states and airlines. Verb semantics are categorized based on their indication of direction. "Directional" verbs such as *'depart'* and *'land'* are distinguished from the others. Prepositions are further split into time-related, direction-related, location-related, cost-related, and other special functions.

4 Experiments

We reimplemented or, in one case (Zhu and Yu, 2017), obtained from the authors code for the models mentioned in Table 4, which also shows the F1-scores reported there. The hyperparameters were set to those that are reported in the papers has having the best performance. Each model was trained for 100 epochs, and then the epoch

Model	Reported F1 score
RNN (Mesnil et al., 2013)	93.98
LSTM (Yao et al., 2014)	95.08
Encoder-Decoder (Kurata et al., 2016)	95.66
Encoder-Decoder with focus (Zhu and Yu, 2017)	95.79
Self-attentive BiLSTM[3] (Li et al., 2018)	96.35
Encoder-Decoder DRL (Wang et al., 2018)	97.86

Table 4: Reported Performance of Models.

with the highest development test set performance was chosen to evaluate on the ATIS test set. We were unable to reproduce comparable figures for the DRL scheme of (Wang et al., 2018) and so it has been excluded from our analysis.

Our own results are reported in Table 5. The column, Test, reports results on the original ATIS test set. Fixed reports on the ATIS test set after all of the repairs mentioned in Section 2.3 were fixed. UNK reports on the ATIS test, with all repairs except the exclusion of utterances with ambiguous occurrences of UNK. Finally, X reports on a corpus, which, similar to the ATIS_X_test set presented in Zhu and Yu (2018), modified the ATIS test set by replacing every NE with a different NE from the same epistemic class in a travel domain ontology defined by them, such that the new NE has never occurred with the same concept. For example, the city *"Toronto"* appears as a *fromloc.city_name* and *toloc.city_name*, but never as a *stoploc.city_name* in ATIS. So *"Toronto"* is used in Corpus X wherever the reference annotation requires a *stoploc.city_name*. Zhu and Yu (2018) did this in order to experiment with a neural architecture that trains first on a coarse classification and then fine-tunes to the ATIS reference annotation in a later step, but the F1 drops on Corpus X are a result of overfitting in which the model effectively learns that Toronto is never a stopover city. Our Corpus X differs from their ATIS_X_test set only in that we first corrected their ontology in light of our taxonomy of annotation errors.

Because the rule-based parser uses an all-paths algorithm, its F1-score is reported in three ways. *Rand(om)* uses the greedy Algorithm 1 in which

Model		Test	Fixed	UNK	X
RNN	Complete	93.56	95.83	94.71	92.3
	Full Parse	93.8	96.8	95.65	93.49
LSTM	Complete	93.86	96.47	95.54	93.29
	Full Parse	94.22	97.44	96.4	94.57
Encoder-Decoder	Complete	94.75	95.77	96.84	91.85
	Full Parse	94.89	96.49	97.55	92.74
Self-att. BiLSTM	Complete	94.87	96.99	96.05	93.60
	Full Parse	95.06	98.02	97.25	94.72
Focus	Complete	95.02	97.61	96.42	84.31
	Full Parse	95.19	98.10	96.86	83.81
Rule-Based[4]	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

Table 5: Experimental Results.

ties are broken at random. *Scep(tical)* only counts successes that every member of a tie produces. *Cred(ulous)* counts successes that any member of a tie produces. The sceptical and credulous scores bracket the possible parse selection strategies. *Full Parse* restricts the evaluation to those utterances (the percentage of which appears in the final row) for which one or more complete parses was found by the rule-based grammar.

5 Analysis and Discussion

One might expect that recent neural approaches could use their word vector representations to generalize better to out-of-domain utterances than the earlier models that Tur et al. (2010) referred to. In fact, the results of the previous section on Corpus X clearly indicate that these recent architectures overfit their language models to filler content itself, overshadowing any potential gain from better contextual inference. ATIS is "shallow" in that it offers only a small amount of training data and an overall lack of lexical and syntactic variety.

What is even more telling is that the performance of these recent architectures on Corpus X is so bad that it falls within the F1 range of our rule-based grammar. The advantages promised by nascent statistical approaches to natural language understanding when rule-based grammars were still in vogue were primarily centred around: (1) portability and (2) coverage. As to portability, recent neural approaches to a corpus as small as ATIS necessarily surrender a certain amount of it for the sake of jointly modeling knowledge of language and domain-specific knowledge — a laudable goal on substantially larger training sets. Our experience with industrial partners suggests, however, that *extensibility*, in which developers wish to roll out the same domain but to a further extent,

[3]The number reported here is with access to the sentence intent labels disabled. In our own runs, reported in Table 5, we disable this model's access to intent labels as well, in order to make a make a more controlled comparison to the other models, none of which use intent labels. Using intent labels, Li et al. (2018) report an F1 score of 96.52%.

[4]The rule-based grammar developer did not have access to the test-domain utterances, and so the grammar replaces OOV test set vocabulary with UNK. These are counted as failures in our statistics unless the UNK token is assigned the correct tag.

such as with more cities, more airports etc. in the case of the ATIS corpus, is of equal importance to them as portability to different domains. There, a rule-based grammar would only be the preferred option if augmenting the filler vocabulary were all that was at stake. It would not be the preferred option if the extension were in the direction of much greater syntactic variety.

That brings us to coverage. The relative error reduction observed after fixing the ATIS annotation generally fails to attain the 50% predicted by Béchet and Raymond (2018). Nevertheless, those repairs put the neural models close to the rule-based grammar's range on utterances for which it generates a full syntactic parse.[5] Our greedy parse selection approach is necessitated by the mere $\sim$80% coverage of the ATIS domain with our rule-based grammar. Neural parsing architectures do exist, and already provide better coverage than 80%.

These arguments taken together suggest that, while there may be very little remaining reward to addressing the slot-filling problem with ATIS, there is still a very perceptible parsing problem, even on a corpus of ATIS's size and lack of syntactic variety. ATIS is not syntactically annotated; to our knowledge, no syntactically annotated corpus in the travel reservation domain exists. The development of such a corpus, the transfer of learning between parsers on different domains of this size, and the appropriation of such a portable parser to slot filling, remain the most promising direction of further research for slot filling, in our view. In this endeavour, ATIS may still play a very prominent role.

References

F. Béchet and C. Raymond. 2018. Is ATIS too shallow to go deeper for benchmarking spoken language understanding models? In *Interspeech*.

B. Carpenter and G. Penn. 1994. The Attribute Logic Engine user's guide, version 2.0. *Laboratory for Computational Linguistics Technical Report, Carnegie Mellon University, Pittsburgh*.

C. T. Hemphill, J. J. Godfrey, and G. R. Doddington. 1990. The ATIS spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

G. Kurata, B. Xiang, B. Zhou, and M. Yu. 2016. Leveraging sentence-level information with encoder LSTM for semantic slot filling. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2077–2083.

C. Li, L. Li, and J. Qi. 2018. A self-attentive model with gate mechanism for spoken language understanding. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3824–3833.

G. Mesnil, X. He, L. Deng, and Y. Bengio. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Interspeech*, pages 3771–3775.

T. Mikolov, S. Kombrink, L. Burget, J. Černockỳ, and S. Khudanpur. 2011. Extensions of recurrent neural network language model. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5528–5531. IEEE.

G. Tur, D. Hakkani-Tür, and L. Heck. 2010. What is left to be understood in ATIS? In *Spoken Language Technology Workshop (SLT), 2010 IEEE*, pages 19–24. IEEE.

Y. Wang, A. Patel, and H. Jin. 2018. A new concept of deep reinforcement learning based augmented general tagging system. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1683–1693.

Y.-Y. Wang, L. Deng, and A. Acero. 2011. Semantic frame-based spoken language understanding. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 41–91.

K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, and Y. Shi. 2014. Spoken language understanding using long short-term memory neural networks. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 189–194. IEEE.

S. Zhu and K. Yu. 2017. Encoder-decoder with focus-mechanism for sequence labelling based spoken language understanding. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5675–5679. IEEE.

S. Zhu and K. Yu. 2018. Concept transfer learning for adaptive language understanding. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 391–399.

[5] Note that on the subset of ATIS test sentences for which our rule-based grammar does obtain a full parse, the neural models also improve, and do attain the predicted 50% RER on the repaired versions of those sentences.