Interspeech 2018
2-6 September 2018, Hyderabad

Is ATIS too shallow to go deeper for benchmarking Spoken Language Understanding models?

Frédéric Béchet[1], *Christian Raymond*[2]

[1]Aix Marseille Univ, Université de Toulon, CNRS, LIS, Marseille, France
[2]INSA Rennes INRIA/IRISA, Rennes, France

frederic.bechet@univ-amu.fr, christian.raymond@irisa.fr

Abstract

The ATIS (Air Travel Information Service) corpus will be soon celebrating its 30th birthday. Designed originally to benchmark spoken language systems, it still represents the most well-known corpus for benchmarking Spoken Language Understanding (SLU) systems. In 2010, in a paper titled "*What is left to be understood in ATIS?*" [1], *Tur et al.* discussed the relevance of this corpus after more than 10 years of research on statistical models for performing SLU tasks. Nowadays, in the Deep Neural Network (DNN) era, ATIS is still used as the main benchmark corpus for evaluating all kinds of DNN models, leading to further improvements, although rather limited, in SLU accuracy compared to previous state-of-the-art models. We propose in this paper to investigate these results obtained on ATIS from a qualitative point of view rather than just a quantitative point of view and answer the two following questions: what kind of qualitative improvement brought DNN models to SLU on the ATIS corpus? Is there anything left, from a qualitative point of view, in the remaining 5% of errors made by current state-of-the-art models?

Index Terms: Spoken Language Understanding, ATIS, Deep Neural Network, Conditionnal Random Fields

1. Introduction

The recent gold rush on Deep Neural Network (DNN) models for handling all kinds of supervised learning tasks on image, audio and text data has increased the need for benchmark datasets that can reliably assess the pros and cons of each new method. One of the most well-known dataset for benchmarking SLU systems is the ATIS (Air Travel Information Service) corpus, containing utterances with semantic annotations corresponding to spoken dialogs about several air travel planning scenarios. Since the introduction of this corpus in 1990, numerous machine learning methods have been applied to the ATIS semantic parsing task with a steady improvement in performance, leading to an accuracy of about 95% at the token level with current neural models such as Bi-LSTM. Since this achievement, more and more complex DNN models have been applied to ATIS leading in some cases to further improvements although rather limited in terms of error reduction.

Looking only at quantitative performance metrics can be justified when dealing with "*big*" datasets and significant variation of error rates. However in the case of ATIS it seems to us that qualitative performance is equally important, especially considering the relatively small size of the dataset and the very limited gains obtained nowadays on SLU accuracy.

In 2010, in a paper titled "*What is left to be understood in ATIS?*", Tur et al. discussed the relevance of this corpus after more than 10 years of research on statistical models for performing SLU tasks, leading to impressive performance ranging from 93 to 95% accuracy with Conditional Random Field (CRF) models. Their conclusion was that although the remaining slice of errors was rather thin, ATIS still contained some ambiguities not yet well covered by the existing models, especially for semantic long distance dependencies, leading to *interesting* errors made because of the limited size of the word contexts handled by SLU models.

The goal of this paper is to give a follow-up to this previous study in the DNN-era. We propose to investigate the latest results obtained on ATIS from a qualitative point of view and answer the two following questions: what kind of qualitative improvement brought DNN models to SLU on the ATIS corpus? Is there anything left, from a qualitative point of view, in the remaining 5% of errors made by current state-of-the-art models?

2. ATIS and Spoken Language Understanding

If large benchmark datasets can be found for tasks such as image and text classification, speech and speaker recognition, this is not the case for tasks such as semantic parsing which combines both the issues of firstly defining formally the annotation model and secondly finding relevant datasets that can illustrate the chosen model. The main difference between Natural Language Understanding (NLU) from text and SLU is the choice of the semantic model: based on a formal linguistic model for NLU (Berkeley Framenet, Abstract Meaning Representation, ...); based on an application for SLU. SLU has been mostly studied in the context of human-machine interaction, such as Spoken Dialog Systems (SDS) for information access or booking service. In this context the semantic model is defined according to the applicative framework and is usually split between a global model which associate a label to a whole utterance (call-type, dialog-act, intent) and a local model in charge of detecting the different concepts or frames occurring in an utterance with possibly semantic links between them. The first issue in obtaining a dataset for the development and evaluation of such an SLU system is the need of a running SDS, leading to a "*chicken and egg*" problem.

One common approach developed to overcome this problem is the simulation of a running system, controlled by a human operator, which is used to collect data from users communicating with it as it was a real automatic system. This *Wizard-of-Oz* method was used in order to collect the DARPA *Air Travel Information Service* (ATIS) corpus [2], containing spoken dialogs about several air travel planning scenarios. Eight research centers participating in the DARPA ATIS project recorded 14,150 utterances corresponding to 1383 scenarios from 398 speakers [3].

10.21437/Interspeech.2018-2256

All utterances were manually transcribed and semantically annotated through an SQL query expressing the meaning of each request with an operational semantic form. The ATIS corpus used nowadays for benchmarking SLU models comes from this original corpus but the semantic annotations have been transferred at the utterance and word levels instead of an SQL query: each utterance is labeled with a global label (often referred as *intent*) corresponding to the 17 kinds of request that exists in the corpus (request about a flight, an airline, meals on board, fare, ...); then each SQL request attribute is projected at the word level through a *Begin, Inside, Outside* (BIO) annotation scheme. We use in this study the widely used version of the ATIS corpus described in [4] where all attributes are collapsed as single words with an additional column containing specific labels for named entities marked via table lookup for city, airline, airport names, and dates. An example of such as corpus is given below on the request: *On april first I need a ticket from Tacoma to San Jose departing before 7am.*

```
on        on          O
april     month_name  B-depart_date.month_name
first     first       B-depart_date.day_number
i         i           O
need      need        O
a         a           O
ticket    ticket      O
from      from        O
tacoma    city_name   B-fromloc.city_name
to        to          O
san-jose  city_name   B-toloc.city_name
departing departing   O
before    before      B-depart_time.time_relative
7am       time        B-depart_time.time
```

We will focus in this paper on attribute prediction, seen as a slot filling task or a sequence labeling task. Considering the example above, the task is to predict the third column knowing the first two. The training corpus contains 4978 request utterances (52K words); the test corpus is made of 893 utterances (8.3K words). The semantic model contains 84 attribute labels. In this study we will use a set of models for performing the slot filling task, each implementing a different machine learning algorithm or paradigm. We will use the output of all systems in order to perform quantitative and qualitative analysis of their results in order to answer the questions addressed in the introduction.

3. SLU models for ATIS slot filling task

According to the literature on sequence labeling tagging [5, 6], 2 conditions are crucial to build robust taggers: using embedding representations as input instead of symbolic ones and model output label dependencies. Recurrent Neural Networks such as Elmann or Jordan models have proven to be very efficient for building taggers on ATIS [7]. Nevertheless, these kinds of models are not very relevant to model output dependencies and remain below *Conditional Random Fields* (CRF) on more complex datasets like MEDIA [5]. Other recurrent models like LSTM or GRU, or eJordan allows better output labels dependencies modeling [6], and CRF can also be used on top of NN [8].

In this paper, we evaluated on ATIS several models with different characteristics: symbolic or numerical input, modeling of target label dependencies or not:

1. **Boost:** we decide to evaluate a boosting algorithm on *bonsai trees* [9]. This algorithm is very relevant as local classifier but it is not dedicated to process sequence to sequence problems, it does not model output labels dependencies and we don't give to him any numerical inputs (word embeddings). This model is expected to be our lowest baseline. We used 1000 bonsai trees of size 2 (4 leaves) on unigrams of token word/relative position.
2. **symCRF:** a standard CRF algorithm with symbolic input features, very relevant to model output labels, with only word features (*symbCRF WO*), or with words+named entities (column 3) (*symbCRF NE*).
3. **neurCRF:** a CRF that uses word embedding input representation.
4. **MLP:** a standard single-hidden-layer feed-forward neural network of size 200; it has the ability to process word embedding representation but does not model any target label dependencies.
5. **BiLSTM:** a bidirectional recurrent LSTM network is used to encode the sequence of word into a vector, followed by a softmax output layer. It implements a 200 (2*100) encoded utterance representation. This system is expected to be the best as processing the observation since it has access to the whole utterance.

All neural based models are build using Keras [10], bonzaiboost implementation has been used for boosting [9] and wapiti [11] for symbolic CRF.

The common parameters for all the experiments are the following: the observation windows is 11 $[-5, +5]$, except for *BiLSTM* and *Boost* that take into account the whole utterance; for comparison with the literature, symCRF was also trained with a windows size of 5 and 2 feature sets: words only (WO) and word/NE (NE); word embeddings are of size 100 and learned jointly with the network; the model selection strategy for neural systems is to keep the best set of parameters among 100 epochs according to the training set; regularization is done using a dropout [12] of 0.5 at the output of the last$_{-1}$ layer of the network.

4. Quantitative evaluation

All five previous models are compared in table 1 using two standard evaluation metrics: F1 computed by conlleval[1] evaluation script that consider a segment correct if both boundaries and class are correct and sclite[2] error rate that do not care about correct boundaries.

model	#params	F1	%error
symbCRF WO$[-2, 2]$	400,950	91.93%	10.1%
symbCRF WO$[-5, 5]$	545,049	92.40%	9.8%
symbCRF NE$[-2, 2]$	130,734	94.30%	6.4%
symbCRF NE$[-5, 5]$	179,739	95.28%	5.4%
boost	255,000	94.96%	5.8%
neurCRF	129,670	95.17%	5.5%
BiLSTM 2*100	210,582	95.30%	5.5%
MLP	269,182	95.74%	5.0%

Table 1: *Systems comparison in terms of F1 computed with conlleval and error rate computed with sclite*

Results presented in table 1 are comparable with previous studies on ATIS [4, 6, 8, 13, 14, 15]. As expected CRF models using only word features obtain the worst results. Adding

[1] https://github.com/tpeng/npchunker/blob/master/conlleval.pl

[2] http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm

named entities helps them generalizing, and makes them comparable to models using word embeddings. Increasing the size of the observation window is also very useful, leading to a big improvement for CRF models. Neural approaches obtain the best results, however it is interesting to notice that recurrent networks don't bring any extra gain compared to a simple MLP with a large observation window (11 words). This can be explained by the small length of sentences in ATIS: 9.3 words on average. Boosting, although not using any embeddings nor sequence features, obtain relatively good performance compare to more sophisticated methods.

Despite some differences in terms of slot prediction performance, it is worth noticing that all these systems give very similar results. As noticed in [13]: "*Based on the number of words in the dataset and assuming independent errors, changes of approximately 0.6% in F1 measure are significant at the 95% level*". Therefore many F1 results published on the corrected ATIS version currently used and released by [4] are not statistically better than the ones reported in the same paper where a 95% F1 score was obtained with SVM and CRF based tagger.

Because of these small differences in performance among systems, percentages can be misleading so it is worth switching to actual numbers to compare systems. Table 2 provides such a comparison. We kept the 5 best systems of table 1 and choose the *symbCRF NE*$[-5, 5]$ setting for *symbCRF*. In table 2 the **#ref** column contains the total number of semantic slots to be found in the ATIS test set; column **#hyp** contains the number of semantic slots hypothesized by each system; column **#ok** presents the number of correct slot hypotheses for each system; column **delta** shows the difference, in terms of correct slot hypotheses, between the best system (**MLP**) and all the others; column **#err** displays the number of erroneous slot hypotheses for each system and finally column **F1** recall the performance of each system in terms of F1 measure.

system	#ref	#hyp	#ok	delta	#err	F1
boost	2800	2832	2674	-25	181	94.96
neurCRF	2800	2834	2681	-18	175	95.17
symbCRF	2800	2830	2682	-17	173	95.28
BiLSTM	2800	2837	2686	-13	173	95.30
MLP	2800	2838	2699	0	159	95.74

Table 2: *Detailed results in numbers of correct and erroneous semantic slot predictions for the 5 systems*

As we can see, the biggest **delta** value is -25, there is only 25 more erroneous decisions at the slot level in the whole corpus between the best and the worst of the 5 systems. The total number of slot errors of the worst system is also rather limited, 181 errors, for a total of 2800 slots to be found. Having 5 systems representing different models (boosting, CRF and DNN) and making different errors can help us characterizing the remaining errors.

The methodology we followed was to partition the slot hypotheses of all systems into 4 clusters: ***AC, NC, AE, NE*** as described in table 3:

- ***AC*** (*Agreement/Correct*) contains the correct slot predictions made by the 5 systems when they all agree (total agreement), therefore it correspond for us to the *solved problem* cluster:
- ***AE*** (*Agreement/Error*) contains the erroneous predictions made by the 5 systems while they all agreed on the same wrong label; this can correspond either to reference errors or to the *open problem* cluster for which we don't have yet a good solution in terms of model prediction;
- ***NC*** (*No agreement/Correct*) correspond to the *comparative* cluster where the 5 systems don't agree on the same label, and at least one prediction is correct; comparison between systems will be made on this cluster;
- ***NE*** (*No agreement/Error*) is another kind of *open problem* cluster since no system found the correct label, although they did not agree on the same wrong one.

cluster	***A (agreement)***	***N (no agreement)***
C (correct)	|AC|=2640	|NC|=79
E (error)	|AE|=124	|NE|=18

Table 3: *Repartition of correct and incorrect slot prediction, by at least one of the 5 systems, when they all agree on a decision, or when there is no agreement*

According to table 3, ATIS is almost a solved problem as cluster ***AC*** contains 2640 of the 2800 slots of the test corpus (94.3%). The amount of erroneous slots being rather limited, we performed a manual inspection of each error in order to present a *qualitative evaluation* of our 5 systems on ATIS. The manual analysis of cluster ***NC*** (*comparative cluster*) was used to address the question: what kind of qualitative improvement brought DNN models w.r.t. previous CRF models? The analysis of the *open problem* clusters ***AE*** and ***NE*** addressed the question: what is not yet modeled by our current systems in ATIS?

5. Qualitative comparison

If we consider the union of all the errors made by the 5 systems on the ATIS test corpus, we obtain a set of 221 wrong predictions, consisting of the union of the 3 clusters NC, AE and NE. We manually checked these 221 semantic slots and found that half of them (110) shouldn't be considered as errors, as they belonged to one of these three categories:

- Errors in the reference labels (51 slots): missing slots, confusion between departure/arrival slots
- Ambiguous sentences (44 slots) where slots could be labeled with different labels. For example in the sentence: "*Show me airlines that have flights between Toronto and Detroit*", *Toronto* is labeled as the departure city and *Detroit* as the arrival city in the reference annotation, however switching arrival and departure labels should not be considered as an error.
- Repetition errors (15 slots). In ATIS, only the first mention of an entity is labeled. Therefore in the sentence: "*Show flight and prices Kansas city to Chicago on next Wednesday arriving in Chicago by 7pm*". In the reference annotation, only the first mention of *Chicago* is labeled as an arrival city, not the second one ("*arriving in Chicago*"), although form a semantic point of view, both mentions should be labeled.

If we remove these 110 mistakes or ambiguous slots from the evaluation process, we obtain the results presented in table 4. As we can see the relative rank of each system is the same as in table 2, however it makes the remaining slice of errors even thinner.

We characterized the remaining 111 errors according to 9 categories as presented in table 5. These categories explain

system	#ref	#hyp	#ok	delta	#err	F1
boost	2742	2738	2666	-25	82	97.30
neurCRF	2742	2740	2674	-17	74	97.56
symbCRF	2742	2736	2677	-14	70	97.74
BiLSTM	2742	2741	2684	-7	65	97.90
MLP	2742	2743	2691	0	59	98.12

Table 4: *Detailed results after removing annotation errors and undecidable ambiguities*

OOV	unknownlocation name orcode
FLIGHT	confusion between start and arrival slots
LIST	unrecognized flight list
DISF	error caused by a disfluency (repetition)
INSERT	slot false detection
GROUND	request for ground transportation confused with flight
MISS-MOD	modifier (location, time) not recognized
RETURN	unrecognized return flight
STOP	unrecognized stop flight

Table 5: *Categorization of semantic slot prediction errors*

the kind of error being made, from the misrecognition of a city name (OOV), the confusion between arrival (date, location, time) and departure in a flight request, or the confusion between a request about ground transportation and flight information.

cluster	***NC***				***AE***
error type	symCRF2	symCRF5	BiLSTM	MLP	all
FLIGHT	**26**	**10**	**11**	**13**	2
LIST	6	5	2	6	4
MISS-MOD	3	3	1	1	2
GROUND	4	5	4	4	**25**
RETURN	3	3	3	3	2
INSERT	1	4	2	1	-
STOP	1	1	1	1	1
DISF	1	1	1	1	2
OOV	3	1	2	2	5
Total	50	33	27	31	43

Table 6: *Distribution of errors for each system according to clusters NC and AE*

Table 6 presents the distribution of errors in the clusters NC and AE. For NC, which is the *comparative cluster* containing slots where at least one system was right and one system was wrong, we show the distribution for 4 systems: two CRF systems differing by the input window size ($[-2, +2]$ for symCRF2, $[-5, +5]$ for symCRF5), the biLSTM and the MLP. For AE, the *open problem* cluster, since all systems produce the same erroneous label for each slot, we provide only one number (column *all*).

It it interesting to see that the error distributions are very different in the two clusters. For AE, more than half the errors come from the single category *GROUND*. This category corresponds to requests about ground transportation that has been misclassified as request for flights, like in the example: *List the distance in miles from New-York's La Guardia airport to downtown New-York city*. In the ATIS corpus, there is no special slot label for this kind of request. Therefore *La Guardia airport* is simply labeled as *airport* although the automatic systems predict the label `fromloc.airport_name`, which is not incorrect from a semantic point of view, but is incorrect in ATIS since this is not a slot about flights. This analysis leads us to think that there is not a specific kind of ambiguity in the ATIS corpus which is not modeled by current models, the dominant ambiguity on the *GROUND* category being more a hole in the semantic annotation scheme than a real source of ambiguity.

For the cluster NC, it is the category *FLIGHT* which is the most dominant one. This is expected as this correspond to the main ambiguity in the ATIS model: recognizing if a given location, date or time is related to the departure or the arrival of a flight request. If this is straightforward in most of the cases, some sentence can be more ambiguous and a large context around the slot to label can be necessary. For example, in the sentence: "What airlines off from Love-Field between 6 and 10 am on June sixth ?", the symCRF2 model predicts the wrong label `arrive_time.end_time` for the slot *10 am*, unlike the other models that correctly predict `depart_time.end_time`.

In our experiments, the symCRF2 method produces twice as many *FLIGHT* errors than the other methods. This shows that the size of the input window is more crucial than the type of model used. Neural methods have the ability to model efficiently large contexts needed to process ambiguous sentences. CRF methods can perform well if they take into account also a large context, although this leads to a big increase in the feature space size. This is acceptable on ATIS because of the low complexity of the semantic model, but it might not be possible on a richer dataset.

The fact that the DNN methods we used did not model output labels dependencies don't seems to be a problem, probably because these methods have access to the whole sequence, each decision is taken according to dependencies spanning over the entire sentence. We have also to point out that there was no issues with position labels (B,I,O) in our experiments since the version of the corpus we used collapsed semantic slots as single tokens. Output label dependencies are much more important when spans of different sizes have to be predicted.

6. Conclusions

This paper proposes a quantitative and a qualitative study of 5 different semantic parsing models on the ATIS SLU slot filling task. From these analyses we were able to answer the two questions cited in the introduction, firstly by noticing that DNN methods perform well for processing ambiguities that need a large context to be removed, with fewer parameters than CRF models with large input windows. However a simple MLP with a larger input window achieved the best results, so it seems that on a simple task as the ATIS SLU task, there is no need for sophisticated models.

To the second question, about the remaining 5% errors in ATIS, we have shown that half the errors were not *real* errors, but rather errors in the reference labels and natural ambiguities, then we did not find in our qualitative analysis any hints that there was a phenomenon non-covered by our models in the very small slice of errors left. To conclude, for us the answer to the question mentioned in the title of this paper is: yes.

7. Acknowledgements

Research supported by ANR16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI), and NVIDIA Corporation with the donation of the GTX Titan X GPU used in this research work.

8. References

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

[2] P. J. Price, "Evaluation of spoken language systems: The atis domain," in *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

[3] L. Hirschman, "Multi-site data collection for a spoken language corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 7–14.

[4] C. Raymond and G. Riccardi, "Generative and discriminative algorithms for spoken language understanding," Antwerp, Belgium, August 2007, pp. 1605–1608.

[5] V. Vukotic, C. Raymond, and G. Gravier, "Is it time to switch to word embedding and recurrent neural networks for spoken language understanding?" in *InterSpeech*, Dresde, Germany, September 2015.

[6] M. Dinarelli, V. Vukotic, and C. Raymond, "Label-dependency coding in Simple Recurrent Networks for Spoken Language Understanding," in *Interspeech*, Stockholm, Sweden, Aug. 2017. [Online]. Available: https://hal.inria.fr/hal-01553830

[7] G. Mesnil, X. He, L. Deng, and Y. Bengio, "Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France, August 25-29, 2013*, 2013, pp. 3771–3775. [Online]. Available: http://www.isca-speech.org/archive/interspeech_2013/i13_3771.html

[8] G. Mesnil, Y. Dauphin, K. Yao, Y. Bengio, L. Deng, D. Hakkani-Tür, X. He, L. Heck, G. Tur, D. Yu, and G. Zweig, "Using recurrent neural networks for slot filling in spoken language understanding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 530–539, 2015.

[9] A. Laurent, N. Camelin, and C. Raymond, "Boosting bonsai trees for efficient features combination : application to speaker role identification," Singapour, September 2014.

[10] F. Chollet *et al.*, "Keras," https://github.com/keras-team/keras, 2015.

[11] T. Lavergne, O. Cappé, and F. Yvon, "Practical very large scale CRFs," in *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, July 2010, pp. 504–513. [Online]. Available: http://www.aclweb.org/anthology/P10-1052

[12] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jan. 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2627435.2670313

[13] K. Yao, B. Peng, Y. Zhang, D. Yu, G. Zweig, and Y. Shi, "Spoken language understanding using long short-term memory neural networks," in *SLT*. IEEE, 2014, pp. 189–194.

[14] B. Liu and I. Lane, "Attention-based recurrent neural network models for joint intent detection and slot filling," in *Interspeech*, 2016.

[15] X. Zhang and H. Wang, "A joint model of intent determination and slot filling for spoken language understanding." in *IJCAI*, 2016, pp. 2993–2999.