LEARNING WITH NOISY SUPERVISION FOR SPOKEN LANGUAGE UNDERSTANDING

Christian Raymond[*]

L.I.A.
University of Avignon, France
christian.raymond@univ-avignon.fr

Giuseppe Riccardi[†]

D.I.S.I.
University of Trento, Italy
riccardi@disi.unitn.it

ABSTRACT

Data-driven Spoken Language Understanding (SLU) systems need semantically annotated data which are expensive, time consuming and prone to human errors. Active learning has been successfully applied to automatic speech recognition and utterance classification. In general, corpora annotation for SLU involves such tasks as sentence segmentation, chunking or frame labeling and predicate-argument annotation. In such cases human annotations are subject to errors increasing with the annotation complexity. We investigate two alternative noise-robust active learning strategies that are either data-intensive or supervision-intensive. The strategies detect likely erroneous examples and improve significantly the SLU performance for a given labeling cost. We apply uncertainty based active learning with conditional random fields on the concept segmentation task for SLU. We perform annotation experiments on two databases, namely ATIS (English) and Media (French). We show that our noise-robust algorithm could improve the accuracy up to 6% (absolute) depending on the noise level and the labeling cost.

***Index Terms*—** Spoken Language Understanding, Active Learning, Conditional Random Fields.

1. INTRODUCTION

Spoken Language Understanding (SLU) aims at extracting concept and their relations from spontaneous speech. We use machine learning algorithms to extract these relations from annotated corpora. The drawback of this approach is the need of a semantically annotated data in order to learn the model. In this context of supervised learning, the two main issues are to reduce the cost of manual annotation and to deal with noisy or mislabeled examples which might impact the statistical learner [1]. The first issue is addressed by the Active Learning (AL) framework which selects for manual annotation the most informative examples and thus reduces the number of supervised training examples needed to achieve a given level of performance. AL has been successfully studied in many tasks: word segmentation, shallow semantic parsing, speech recognition and spoken language understanding [2]. The second issue is annotation error detection and control. The main idea is to detect (likely) noisy annotations and improve their accuracy [3, 4]. Most of the work related to labeling error detection deals with the problem as a post-processing step but it is an important part of the labeling process [5, 6].

This paper presents an uncertainty based AL framework in order to deal with annotation errors. We focus on the concept segmentation task for SLU using CRFs [7] which have been shown to be efficient [1]. We use the probabilistic confidence of the CRF model [8] to assign the degree of uncertainty to a whole annotation. We propose two noise-robust strategies based on error detection. One is data-intensive since no more human supervision is needed, the total amount of annotated data is not affected while the strategy improves the training data set significantly in comparison of the standard AL. The second is supervision-intensive and proposes for human correction likely erroneous annotations. This strategy reduces the total amount of annotated data but exhibits better performance on the model side in comparison to the previous one for a given labeling effort. We show on two SLU databases, namely ATIS and MEDIA, that the noise-robust strategies could improve the accuracy up to 6% (absolute) depending on the noise level and the labeling cost.

The paper is structured as follow. We present the two SLU datasets used in our experiments in the section 2. The section 3 presents the adapted strategies to noisy datasets and compares them against passive learning and traditional active learning.

2. MEDIA AND ATIS DATASETS

In our experiments we used two datasets. ATIS [9] is a publicly available corpus used in the early nineties for SLU evaluation. MEDIA [10] has been recently collected and will be made available through ELRA.

ATIS: The Air Travel Information System (ATIS) task is designed to provide flight information. The semantic representation used is frame based. The SLU goal is to map the language query into a frame/slot structure. In this paper we focus

[*] The author performed the work while he was at the university of Trento

[†] This work has been partially funded by the European Commission - LUNA project contract n^o 33549 and by the Marie Curie Excellence Grant for the ADAMACH project (contract n^o 022593).

on the extraction of frame slots represented as attribute/value pair (concept). For example, from the user request "list flights from boston to philadelphia", the concept "*FROMLOC.CITY = boston*" and "*TOLOC.CITY = philadelphia*" have to be extracted. We start from the same dataset as [11]: the training set consists of 4978 utterances selected from the Class A (context independent) training data in the ATIS-2 and ATIS-3 corpora whilst the ATIS test set contains both the ATIS-3 NOV93 and DEC94 datasets, 893 utterances. In [11], each training utterance is annotated with an abstract semantic annotation generated from the hand corrected semantic parse results from the Phoenix parser. We did a semi-automatic procedure to get the words/concept alignment.

MEDIA: The French research project MEDIA evaluates different SLU spoken dialogue systems designed to provide tourist information. The dialog corpus includes 1250 Wizard of Oz conversation recordings: 250 speakers have followed each 5 hotel reservation scenarios. This corpus has been manually transcribed and then conceptually annotated according to a semantic representation defined within the project. This representation is based on the definition of concepts that can be associated to 3 kinds of information: the standard attribute/value information, specifier information, and mode (see [10] for more details). Table 1 shows an example of message from the MEDIA corpus with the attribute/value information. The semantic dictionary MEDIA contains 83 concept labels, 19 specifiers and 4 modal information. In this study we focus on the concept extraction only. No specifiers, values or modal information are considered. The MEDIA corpus is split into 3 parts. The first part (720 dialogues, 12K messages) is used to train the models, the second (79 dialogues, 1.3K messages) and the third part (200 dialogues, 3.4K messages) are used as test.

Table 1. Example of the semantic attribute/value (without specifiers) representation for the sentence "*yes the hotel whose price doesn't exceed one hundred and ten euros*".

word seq.	attribute name	attribute value
oui	response	oui
l'	refLink-coRef	singulier
hôtel	BDObject	hotel
dont	null	
le prix	object	paiement-montant
ne dépasse pas	comparative-payment	inferieur
cent dix	payment-amount-integer	110
euros	payment-unit	euro

3. NOISE-ROBUST AL STRATEGIES

Data-driven SLU systems need semantically annotated data which are expensive, time consuming and prone to human errors. AL aims to minimize the number of labeled utterances by automatically selecting for labeling the utterances that are likely to be most informative. But annotation errors still strongly impact the statistical systems performances. Annotation consistency and reliability is thus a crucial issue. We expect that integrating an error detection algorithm in the AL process will be beneficial for the following reasons: detecting annotation errors is beneficial to improve the train set quality in order to learn accurate model; it exhibits ambiguous annotations and could improve the annotator skill at each AL turn.

The idea behind the error detection algorithm is that the examples which are classified with low confidence with a model trained with the very same data are more probably labeling errors or outliers. Distinguish between errors or hard examples is not trivial, and error detection methods do not achieve very good precision and recall at the same time [4, 12]. The intuition is that an example is hard to learn because the feature(s) needed to discriminate it from others is(are) not present(s). Our experiments confirm that some concepts confusion pairs persist during all the AL turns: the learner repeats the same mistake. It means that adding more examples does not help the learner. Obviously, removing erroneous examples will be benefit, we guess that removing hard ones will not have impact on the learner performance. Moreover, checking hard examples could be useful to find the features which should be incorporated into the model, which are in some cases very intuitive [1]. According to these facts, we propose two noise-robust strategies plugged into the AL framework to deal with annotation errors during the annotation process.

We experiment these strategies within the certainty-based AL method which selects for labeling the examples that the learner is least confident about. The learner used is using Conditional Random Fields (CRF) [7] which provides a conditional probability model over the whole annotation given the observations. As [8] we exploit this probability to measure uncertainty. We compare the strategies with the passive approach where the new examples to be labeled are chosen randomly and the standard AL algorithm.

The figure 1 presents the AL algorithms. The standard AL follows the steps 1,2(a,b,c,d,h,i) while the noise robust strategies execute the steps 2(e,f,g) too. The standard AL starts (step 1) with N bootstrap randomly annotated examples S_L to build a first model μ_0, N is about 10% of available transcription (*i.e.* 500 examples for ATIS, 1000 for MEDIA). μ is a CRF trained using a traditional first order dependency graph. Features are the indicators for specific words and their corresponding lexical class in a window [-4, 2] around the decision state. In each AL turn, a batch of k examples for which the model μ is less confident about is selected (S_k) in the unlabeled part S_U. Then S_k is presented for human labeling and added to the set of training data. A new model μ is trained and the process is repeated. We use a batch selection $k = 200$ for ATIS and $k = 1000$ for MEDIA. This is our baseline B.

The bold part (steps 2(e,f,g)) in figure 1 correspond to the standard algorithm modification and is followed by the noise-robust strategies:

1) the first strategy, following the step 2(g)i in figure 1, is

```
1. Train a model $\mu$ using small amount of $N$ labeled data randomly selected ($S_L$)
2. while (labeler/data available)
    (a) Use $\mu$ to automatically label the unlabeled part of the corpus ($S_U$)
    (b) Rank automatically annotated examples ($S_U$) according to the confidence measure given by $\mu$
    (c) Select a batch of $k$ examples with the lowest score ($S_k$)
    (d) Ask for human labeling on $S_k$
    ------------------------------------------------------------
    (e) Use $\mu$ to automatically label $S_L$
    (f) Rank automatically annotated examples ($S_L$) according to the confidence measure given by $\mu$
    (g) select $I$ examples under a given threshold ($S_I$)
        i. STRATEGY 1: remove ($S_I$) from the train data ($S_L = S_L - S_I$)
        ii. STRATEGY 2: Manually check/correct $S_I$ to obtain $S'_I$ then ($S_L = S_L - S_I + S'_I$)
    ------------------------------------------------------------
    (h) $S_L = S_L + S_k$
    (i) Train a new model $\mu$ with $S_L$
```

Fig. 1. Active Learning algorithm: the standard AL follows the steps 1,2(a,b,c,d,h,i), the noise-robust AL algorithms follow the baseline + the steps 2(e,f,g) in bold

data-intensive: at each turn of the AL algorithm, detects and removes automatically the I likely erroneous annotated examples from the training set and thus increases the accuracy of the model. The strategy does not ask for further human supervision for each data sample.

2) the second, following the step 2(g)ii, is *supervision intensive*: the strategy selects, at each AL turn, the I most likely erroneous annotated examples and asks further human supervision. Both the model and the human annotators should benefit by this strategy: on the model side, the train set is cleaner and more consistent, on the annotators side, some cases of annotation errors are disambiguated and should not occur any more in the future annotations. This strategy privileges the refinement of per-sample annotation accuracy.

We experiment these strategies on noisy ATIS and MEDIA corpora. We use the ATIS dataset presented in the section 2 since it contains annotation errors. In order to perform similar experiments on MEDIA, we produce noisy versions of the MEDIA dataset following the next mechanism:

1) we compute the concept confusion pair statistics from the test set with a model built on all training data,

2) we choose randomly J annotated examples from the MEDIA data,

3) for each chunks of these examples, we decide randomly (0 or 1) if the chunks will be corrupted, if yes, we assign randomly a new concept from the list of confusable concepts.

We produced two MEDIA noisy corpus with different levels of noise varying the parameter J: the first one contains about 2000 corrupted sentences and 13% of concepts corrupted in comparison with the original one, the second has 4000 sentences and 26% of concepts corrupted. In order to simulate the human supervision proposed in the second strategy, we use as checked/corrected annotation, for MEDIA, annotations contained in the original corpus. For ATIS, we disambiguated most of the annotation errors with a simple rule [1] to produce an ATIS *disambiguated* dataset (10% of the annotations have been modified). This dataset is used as the dataset which contains the checked version of annotations.

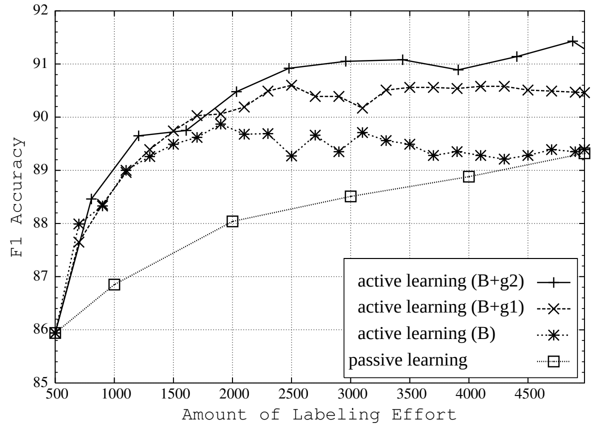

Fig. 2. Noise-robust strategies vs. baseline AL for ATIS

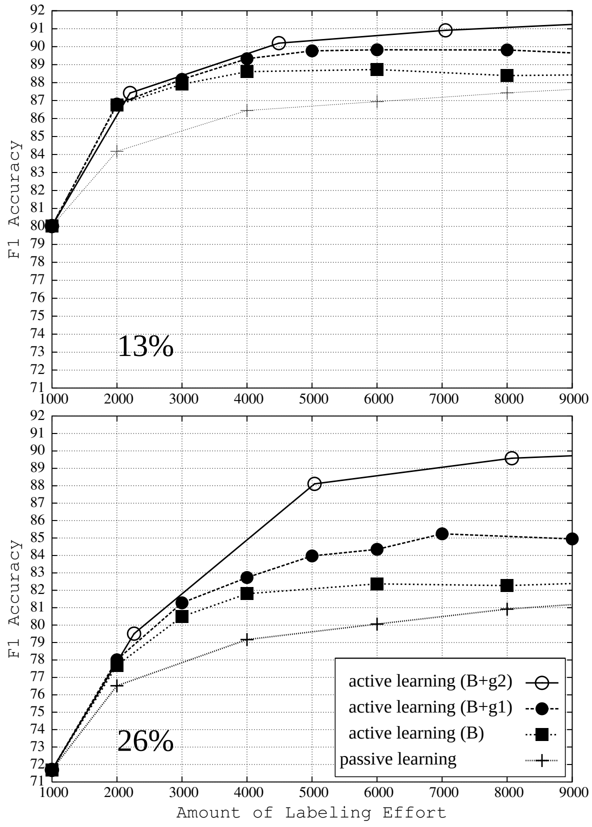

Fig. 3. Noise-robust strategies vs. baseline AL for MEDIA for different noise levels, 13% and 26% of concepts corrupted

To evaluate our strategies we speak in terms of "amount of labeling effort". The first proposed strategy removes some instances at each step of the AL algorithm, thus the model is built with a smaller number of instances but the amount of labeling effort is identical as the standard AL since the same amount of data is annotated and no more human supervision is needed. The second strategy presents to human annotators for checking suspected erroneous examples, the amount of labeling effort is then increased by the size of checked data. This might not be representative of the actual cost because the task of checking and correcting instances differs from annotating them from scratch.

The figures 2 and 3 present the two strategies integrated with the AL baseline B for both ATIS and MEDIA, the data-intensive strategy is denoted "active learning (B+g1)" and the supervision-intensive "active learning (B+g2)". In real condition, I in the step 2g could be parameterized according to the expected level of noise in the annotations (according to the annotation complexity, the number of annotators, their skill, *etc.*). I could be fixed high at the beginning of the annotation process, since the number of annotation errors is expected higher, and decreased each AL turn. In case of the supervision-intensive strategy, I can be adapted to the annotators capabilities and the wanted ratio quality/amount of annotations. The threshold himself in the step 2g should be fixed according to the task complexity. Simpler is the task, more confident should be the model on the clean training data. Anyway this parameter is not constraining since strategies work well with different configurations. Figures 2 and 3 present performance of strategies selecting all annotations under a threshold in step 2g of the algorithm. The threshold determined empirically is 0.7 for ATIS and 0.4 for MEDIA. We can see that noise impact strongly on the learner performance: in order of comparison the accuracy upper bound obtained with the clean version of corpora using the same learner is 95% for ATIS and 92% for MEDIA. The standard AL in situation of noisy data both reduce the annotation cost and improve the upper bound accuracy obtained with the passive learning approach. The data and supervision intensive strategies both improve this AL baseline, up to 3% and 6% respectively depending on the noise level and the annotation effort. The second strategy focus on the reliability of a smaller set of annotated data, and thus more adapted if the goal is to build accurate model as fast as possible. The first strategy does not affect the total amount of annotated data, this could be an important criterion if the data have to be re-used for other purpose.

4. CONCLUSION

We proposed in this paper two strategies to cope with annotation errors in the active learning framework. The idea is based on annotation error detection at each active learning turn. The first one removes automatically these errors from the training set and compute models with higher accuracy in comparison to a standard active learning procedure without affecting the amount of labeled data. The second one asks for human annotators to check them, the total amount of labeled data is reduced but this strategy improves the accuracy in comparison to the first one for an equivalent annotation cost.

Acknowledgment

We would like to thank Yulan He for sharing with us her ATIS annotated dataset.

5. REFERENCES

[1] C. Raymond and G. Riccardi, "Generative and discriminative algorithms for spoken language understanding," in *Interspeech*, 2007.

[2] D. Hakkani-Tür, G. Riccardi, and G. Tur, "An active approach to spoken language processing," *ACM Trans. Speech Lang. Process.*, vol. 3, no. 3, pp. 1–31, 2006.

[3] S. Abney, R. Schapire, and Y. Singer, "Boosting applied to tagging and PP attachment," in *EMNLP/VLC*, 1999.

[4] T. Nakagawa and Y. Matsumoto, "Detecting errors in corpora using support vector machines," in *ACL*, 2002.

[5] A. Vlachos, "Active Annotation," in *EACL*, 2006.

[6] G. Tur, M. Rahim, and D. Hakkani-Tür, "Active labeling for spoken language understanding," in *Eurospeech*, 2003.

[7] J. Lafferty, A. Mccallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, 2001, pp. 282–289.

[8] C. Symons, N. Samatova, R. Krishnamurthy, B. Park, T. Umar, D. Buttler, T. Critchlow, and D. Hysom, "Multi-criterion active learning in conditional random fields," in *ICTAI*, 2006, pp. 323–331.

[9] D. Dahl, M. Bates, M. Brown, W. Fisher, K. Hunicke-Smith, D. Pallett, C. Pao, A. Rudnicky, and E. Shriberg, "Expanding the scope of the atis task: the atis-3 corpus," in *HLT*, 1994, pp. 43–48.

[10] H. Bonneau-Maynard, S. Rosset, C. Ayache, A. Kuhn, and D. Mostefa, "Semantic annotation of the french media dialog corpus," in *InterSpeech*, 2005.

[11] Y. He and S. Young, "Semantic processing using the hidden vector state model," *Computer Speech and Language*, vol. 19, no. 1, pp. 85–106, 2005.

[12] E. Eskin, "Detecting errors within a corpus using anomaly detection," in *NAACL*, 2000, pp. 148–153.