ASR error management for improving spoken language understanding
Edwin Simonnet, Sahar Ghannay, Nathalie Camelin, Yannick Estève, Renato de Mori

To cite this version:
Edwin Simonnet, Sahar Ghannay, Nathalie Camelin, Yannick Estève, Renato de Mori. ASR error management for improving spoken language understanding. Interspeech 2017, Aug 2017, Stockholm, Sweden. hal-01526298

HAL Id: hal-01526298
https://hal.science/hal-01526298
Submitted on 23 May 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
ASR error management for improving spoken language understanding

Edwin Simonnet¹, Sahar Ghannay¹, Nathalie Camelin¹, Yannick Estève¹, Renato De Mori²³

¹LIUM, Le Mans University, France
²LIA, University of Avignon, France
³McGill University, Montreal, Canada

firstname.lastname@univ-lemans.fr, rdemori@cs.mcgill.ca

Abstract

This paper addresses the problem of automatic speech recognition (ASR) error detection and their use for improving spoken language understanding (SLU) systems. In this study, the SLU task consists in automatically extracting, from ASR transcriptions, semantic concepts and concept/values pairs in a e.g touristic information system. An approach is proposed for enriching the set of semantic labels with error specific labels and by using a recently proposed neural approach based on word embeddings to compute well calibrated ASR confidence measures. Experimental results are reported showing that it is possible to decrease significantly the Concept/Value Error Rate with a state of the art system, outperforming previously published results performance on the same experimental data. It also shown that combining an SLU approach based on conditional random fields with a neural encoder/decoder attention based architecture, it is possible to effectively identifying confidence islands and uncertain semantic output segments useful for deciding appropriate error handling actions by the dialogue manager strategy.

Index Terms: spoken language understanding, speech recognition, robustness to ASR errors

1. Introduction

In spite of impressive research efforts and recent results, systems for semantic interpretation of text and speech still make errors. Some of the problems common to text and speech are: difficulty of concept mention localization, ambiguities intrinsic in localized mentions, deficiency to identify sufficient contextual constraints for solving interpretation ambiguities. Additional problems are introduced by the interaction between a spoken language understanding (SLU) system and an error prone automatic speech recognition (ASR) system. ASR errors may affect the mention of a concept, the value of a concept instance. Furthermore, the hypothesization of concepts and values depends, among other things, on the context in which their mention is localized. Thus, context errors may also introduce errors in concept mention location and hypothesization.

The focus of this paper¹ is on the introduction of suitable ASR confidence measures for localizing ASR word errors that may affect SLU performance. They are used as additional SLU features to be combined with lexical and syntactic features useful for characterizing concept mentions. For this purpose, an ASR error detection sub-system has been endowed with confidence features based on syntactic dependencies and other semantically relevant word features. Two SLU architectures equipped with other sets of confidence and specific word features are introduced. The architectures are based on conditional random fields (CRF) and an encoder-decoder neural network structure with a mechanism of attention (NN-EDA). Experimental results showing significant reduction for the French MEDIA corpus on concepts and concept value pairs confirm the expected benefit of introducing semantic specific ASR features.

2. Related work

SLU systems are error prone. Part of them are caused by certain types of ASR errors. In general, ASR errors are reduced by estimating model parameters by minimizing the expected word error rate [2]. The effect of word errors can be controlled by associating a single sentence hypothesis with word confidence measures. In [3] methods are proposed for constructing confidence features for improving the quality of a semantic confidence measure. Methods proposed for confidence calibration are based on the maximum entropy model with distribution constraints, the conventional artificial neural network, and the deep belief network (DBN). The latter two methods show slightly superior performance but higher computational complexity compared to the first one. More recently [4], new features and bidirectional recurrent neural networks (RNN) have been proposed for ASR error detection. Most SLU systems reviewed in [5] generate hypotheses of semantic frame slot tags expressed in a spoken sentence analyzed by an ASR system. The use of deep neural networks (DNN) appeared in more recent systems as described in [6]. Bidirectional RNNs with long-short term memory (LSTM) have been used for semantic frame slot tagging [7]. In [8], LSTMs have been proposed with a mechanism of attention for parsing text sentences to logical forms. Following [9], in [10] a convolutional neural network (CNN) is proposed for encoding the representation of knowledge expressed in a spoken sentence. This encoding is used as an attention mechanism for constraining the hypothesization of slot tags expressed in the same sentence. Most recent papers using sophisticated SLU architectures based on RNNs have the best sequence of word hypotheses as input passed by an ASR system. In this paper, two SLU architectures are considered. The first one, based an encoder with bidirectional gated recurrent units (GRU) used for machine translation [11], integrates context information with an attention based decoder as in [12]. The second one integrates context information in the same architecture used in [1] based on conditional random fields (CRF). Both SLU systems receive word hypotheses generated by the same ASR sub-system and scored with confidence measures computed by a neural archi-

¹Thanks to the ANR agency for funding through the CHIST-ERA ERA-Net JOKER under the contract number ANR-13-CHR2-0003-05.
3. ASR error detection and confidence measure

Two different confidence measures are used for error detection. The first one is the word posterior probability computed with confusion networks as described in [2]. The other one is a variant of a new approach, introduced in [13, 14]. The latter measure is computed with a Multi-Stream Multi-Layer Perceptron (MS-MLP) architecture, fed by different heterogeneous confidence features. Among them, the most relevant for SLU are word embeddings of the targeted word and its neighbors, length of the current word, language model backoff behavior, part of speech (POS) tags, syntactic dependency labels and word governors. Other features, such as prosodic features and acoustic word embeddings described in [15] and [14] could also be used but were not considered in the experiments described in this paper. A particular attention was carried on the word embeddings computation, which is the result of a combination of different well known word embeddings (CBOW, Skip-gram, GloVe) made through the use of a neural auto-encoder in order to improve the performances of this ASR error detection system [16].

The MS-MLP proposed here for ASR error detection has two output units. They compute scores for Correct and Error labels associated with an ASR generated hypothesis. This hypothesis is evaluated by the softmax value of the labels associated with an ASR generated hypothesis. This hypothesis is the result of a combination of different well known word embeddings (CBOW, Skip-gram, GloVe) made through the use of a neural auto-encoder in order to improve the performances of this ASR error detection system [16].

The MS-MLP proposed here for ASR error detection has two output units. They compute scores for Correct and Error labels associated with an ASR generated hypothesis. This hypothesis is evaluated by the softmax value of the labels associated with an ASR generated hypothesis. This hypothesis is the result of a combination of different well known word embeddings (CBOW, Skip-gram, GloVe) made through the use of a neural auto-encoder in order to improve the performances of this ASR error detection system [16].

The MS-MLP proposed here for ASR error detection has two output units. They compute scores for Correct and Error labels associated with an ASR generated hypothesis. This hypothesis is evaluated by the softmax value of the labels associated with an ASR generated hypothesis. This hypothesis is the result of a combination of different well known word embeddings (CBOW, Skip-gram, GloVe) made through the use of a neural auto-encoder in order to improve the performances of this ASR error detection system [16].

The MS-MLP proposed here for ASR error detection has two output units. They compute scores for Correct and Error labels associated with an ASR generated hypothesis. This hypothesis is evaluated by the softmax value of the labels associated with an ASR generated hypothesis. This hypothesis is the result of a combination of different well known word embeddings (CBOW, Skip-gram, GloVe) made through the use of a neural auto-encoder in order to improve the performances of this ASR error detection system [16].
account a weighted sum of all the annotations computed by the encoder. This weighting depends on the current output target, and is the core of the attention mechanism: a good estimation of these weights $\alpha_{ij}$ allows the decoder to choose parts of the input sequence to pay attention to. This context vector is used by the decoder in conjunction with the previous emitted label output $y_{t-1}$ and the current state $s_t$ of the hidden layer of a RNN to make a decision about the current label output $y_t$. A more detailed description of recurrent neural networks and attention based ones can be found in [11].

$$\alpha_{ij} = \frac{\exp(v_j)}{\sum_k \exp(v_k)}$$

![Figure 2: Architecture of the NN-EDA SLU system.](https://gforge.inria.fr/projects/discretize4crf/)

4.3. CRF system

Past experiments described in [1] have shown that the best semantic annotation performance on manual and automatic transcriptions of the MEDIA corpus were obtained with CRF systems. More recently in [19], this architecture has been compared to popular bi-directional RNN (bi-RNN). The results were that CRF systems outperform a bi-RNN architecture on the MEDIA corpus, while better results were observed by bi-RNN on the ATIS [20] corpus. This is probably explained by the fact that MEDIA contains semantic contents whose mentions are more difficult to disambiguate and CRFs make it possible to exploit complex contexts more effectively.

For the sake of comparison with the best SLU system proposed in [1], the Wapiti toolkit was used [21]. Nevertheless, the set of input features used by the system proposed in this paper is different from the one used in [1]. Among the novelties used in our system, we consider syntactic and ASR confidence features and our configuration template is different. After many experiments performed on DEV, our final feature template includes the previous and following instances for words and POS in a unigram or a bigram to associate a semantic label with the current word. Also associated with the current word are semantic categories of the two previous and two following instances. The other features are only considered at the current position.

Furthermore, the tool discretize4CRF is used to apply a discretization function to the ASR confidence measures in order to obtain several discrete values that can be accepted as input features by the CRF.

5. Experimental setup and results

Experiments were carried with the MEDIA corpus as in [1]. For the sake of comparison, the results of their best understanding were performed on DEV, our final feature template includes the previous and following instances for words and POS in a unigram or a bigram to associate a semantic label with the current word. Also associated with the current word are semantic categories of the two previous and two following instances. The other features are only considered at the current position.

Furthermore, the tool discretize4CRF is used to apply a discretization function to the ASR confidence measures in order to obtain several discrete values that can be accepted as input features by the CRF.

system is reported in this paper as baseline. However, as the WER of the ASR used in this paper is lower (23.5%) than the one used in the baseline, rigorous conclusions can be drawn only on comparisons between the different SLU components introduced in this paper.

5.1. The MEDIA corpus

The MEDIA corpus was collected in the French Media/Evalda project [22] and deals with negotiation of tourist services. It contains three sets of telephone human/computer dialogues, namely: a training set (TRAIN) with approximately 17.7k sentences, a development set (DEV) with 1.3k sentences and an evaluation set (TEST) containing 3.5k sentences. The corpus was manually annotated with semantic concepts characterized by a label and its value. Other types of semantic annotations (such as mode or specifiers) are not considered in this paper to be consistent with the experimental results provided in [1]. Annotations also associate a word sequence to the concepts. These sequences have to be considered as estimations of concept localized mentions. Evaluations are performed with the DEV and TEST sets and report concept error rates (CER) for concept labels only and concept-value error rates (CVER) for concept-value pairs. It is worth mentioning that the number of concepts annotated in a turn has a large variability and may include more than 30 annotated concepts. Among the concepts types there are some, such as three different types of REFERENCE and CONNECTOR of application domain entities. The mentions of these concepts are often short confusable sequences of words.

5.2. LIUM ASR system dedicated to MEDIA

For these experiments, a variant of the ASR system developed by LIUM that won the last evaluation campaign on French language has been used [23]. This system is based on the Kaldi speech recognition toolkit [24]. The training set used to estimate the DNN (Deep Neural Networks) acoustic models parameters consists of 145,781 speech segments from several sources: the radiophonic broadcast ESTER [25] and ESTER2 [26] corpora, which accounts for about 100 hours of speech each; the TV broadcast ETAPE corpus [27], accounting for about 30 hours of speech; the TV broadcast REPERE train corpus, accounting for about 35 hours of speech and other LIUM radio and TV broadcast data for about 300 hours of speech. As a total, 565 hours of speech composes the training corpus. These recordings were converted to 8kHz before training the acoustic models in order to be more appropriate to the MEDIA telephone data. As inputs, DNN are fed (for training and decoding) with MFCCs (Mel-Frequency Cepstrum Coefficients) concatenated to i-vectors, in order to adapt acoustic models to speakers.

The vocabulary of the ASR system contains all the words present in the MEDIA training and development corpora, so about 2.5K words. A first bigram language model (LM) is applied during the decoding process to generate word-lattices. These lattices are then rescored by applying a 3-gram language model. In order to get an SLU training corpus close to the test corpus, SLU models are trained with ASR transcriptions. To avoid to deal with errors made by an LM over-trained on the MEDIA training corpus, a leave-one-out approach was followed: all the dialogue files in the training and the development corpora were randomly split into 4 subsets. Each subset was transcribed by using an LM trained on the manual transcriptions present in the 3 other blocks and linearly interpolated to a ‘generic’ language model trained on a large set of French newspapers crawled on the web, containing 77 millions of words.
The test data was transcribed with an LM trained on the MEDIA training corpus and the same generic language model. As shown in Table 2, word error rates for the training, development, and test corpora were around 23.5%.

<table>
<thead>
<tr>
<th></th>
<th>train.</th>
<th>dev.</th>
<th>test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>23.7%</td>
<td>23.4%</td>
<td>23.6%</td>
</tr>
</tbody>
</table>

Table 2: Word error rates of transcriptions produced by the ASR systems in order to train, to tune and to test the SLU systems.

### 5.3. Results

Tests were performed for both architectures with the MEDIA DEV set. The best configuration is chosen with respect to the best results observed on the DEV set and applied for obtaining the TEST results. These results in terms of error rate, precision and recall for concepts (C) and concept value (CV) are reported for the best configuration of each architecture in Table 3.

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>23.8</td>
<td>-</td>
</tr>
<tr>
<td>NN EDA</td>
<td>22.3</td>
<td>0.88</td>
</tr>
<tr>
<td>CRF</td>
<td>19.9</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table 3: Comparison on error rate, precision and recall for concepts (C) and concept value (CV) pairs obtained on TEST.

It appears that the CRF architecture significantly outperforms NN EDA that showed minor improvements with respect to the baseline.

In order to evaluate the impact of the use of confidence measures among the input features, we made some experiments summarized in Table 4. As we can see, the confidence measure provided by the MS-MLP architecture brings relevant information to reduce the CER and the CVER.

<table>
<thead>
<tr>
<th></th>
<th>C</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>baseline</td>
<td>23.8</td>
<td>-</td>
</tr>
<tr>
<td>NN EDA</td>
<td>22.3</td>
<td>0.88</td>
</tr>
<tr>
<td>CRF</td>
<td>19.9</td>
<td>0.90</td>
</tr>
</tbody>
</table>

Table 4: Impact of the use of confidence measures (probability a posteriori and MS-MLP values) on the performances of CRF on TEST.

Other versions of the two systems were considered by adding to the usual MEDIA concept labels two more output tags. During training, these tags are replacing the usual one when the hypothesized word is erroneous. If the erroneous hypothesized word is supporting a concept, it is associated to the ERROR-C tag, ERROR-N otherwise. During evaluation, ERROR-C and ERROR-N hypothesized tags are replaced by null (tag informing that the word does not convey any MEDIA information) in order to perform the usual MEDIA evaluation protocol. Results on TEST, obtained with the best configuration observed on DEV, are reported in Table 5.

### 6. Conclusions

Two variations of two SLU architectures respectively based on CRFs and NN-EDA have been considered. Using the MEDIA corpus, they were compared with the CRF SLU, considered as baseline that provided the best results among seven different approaches as reported in [1]. The main novelties of the proposed SLU architectures are the use, among others, of semantically relevant confidence and input features. The CRF architectures outperformed the NN-EDA architectures with significant improvement over the baseline. Nevertheless, NN-EDA architectures appeared to be useful when combined with the CRF ones. The results show that the interaction between the ASR and SLU components is beneficial. Furthermore, all the architectures show that most of the errors are for concepts whose mentions are made of short confusable sequences of words that remain ambiguous even if they can be localized. These concept types are difficult to detect, even on manual transcription, indicating that the interpretation of the MEDIA corpus is particularly difficult. Thus, suggested directions for future work should consider new structured mechanisms of attention capable of selecting features of distant contexts in a conversation history. The objective is to identify a sufficient set of context features for disambiguating local concept mentions.
7. References


