RECOGNITION ERROR PROCESSING FOR SPEECH UNDERSTANDING

Caroline Bousquet-Vernhettes, Nadine Vigouroux

IRIT UMR CNRS 5505
Université Paul Sabatier
118 Route de Narbonne
31062 Toulouse, France
(bousquet, vigourou)@irit.fr

ABSTRACT

The aim of this paper is to propose an extension of the stochastic conceptual model in order to increase the robustness of the understanding process faced with misrecognitions and unknown words. Corpus analysis shows that some misrecognized words are more difficult to interpret than others, so we defined a word ambiguity rate. We performed trial series on train schedule inquiry application to evaluate the understanding rate when faced with misrecognized words and in particular, when these words are city names.

1. INTRODUCTION

Conversational automatic inquiry systems must allow a large public to express themselves in a spontaneous way. The speech understanding process must be able to understand a large public in spite of the difficulties that are inherent to spoken natural language: ambiguities, variability and choices of expression (indeed users express themselves spontaneously, so they can hesitate, make errors or correct themselves) and errors introduced by the automatic speech recognition itself.

The aim of this paper is to propose an approach for robust understanding face with recognition errors.

The two main recognition errors are due to 1) use conditions very different than training conditions of acoustic models (noisy environment…) and 2) the use of unknown words (words that are not covered by the recognition system’s vocabulary, these words are also called out-of-vocabulary words). The pronounced word is usually confused with the most similar sounding word (or word group) present in the lexicon.

The presence of a misrecognized word in the speech understanding module input causes semantic ambiguity: it is necessary to detect that this word is wrong and then interpret it. In the same way, the meaning of an unknown word must be identified.

Some authors tried to automatically detect and label OOV words at the speech recognition level, like in [1] and [2]. Some speech understanding systems ([3], [4]) are able to interpret an utterance including a word labeled OOV by the speech recognition. The classical approach to deal with misrecognized words consists on determining if a word is correct or not, using confidence score given by the recognizer (see for instance [5]).

One of the challenges is to increase the robustness of speech understanding face with recognition errors. So, the speech understanding process must be able to interpret words labeled as OOV words by the speech recognition, and to detect if a word is incorrectly recognized and interpret it.

In this paper, we present a new approach for the integration of misrecognitions and OOV words into stochastic language models to increase the robustness of the understanding process.

We performed trial series on train schedule inquiry application to evaluate the performances of our model when it is faced with misrecognized words and in particular city names. We will see that interpretation of misrecognitions is difficult and the results depend on the misrecognition type. So, we define a word ambiguity rate to characterize this difficulty.

In a first section, we present the conceptual and stochastic approach and then, how this basic model is extended for the processing of misrecognized words. The second section details the applicative framework (train schedule application). Then, results obtained from recognition output are given and discussed. Finally, the last section presents a test about the interpretation of misrecognized towns.

2. STOCHASTIC CONCEPTUAL MODEL

2.1. Conceptual segment approach

We have developed an understanding system based on stochastic representation of semantic and pragmatic knowledge using conceptual segments. This approach is inspired to the CHRONUS method [6].

A conceptual segment is a word sequence corresponding to the basic units of meaning. We distinguish three kinds of conceptual segments: referential representing the application domain, illocutionary referring to the speech act theory and filler. A filler conceptual segment is used to model all words or word sequences judged as irrelevant for the meaning representation. A word sequence labeled as filler usually corresponds to user’s digressions or extra-linguistic phenomena.

In task-oriented domain (like train or airline reservations, cinema schedules and so on), the set of conceptual segments is finite. Conceptual segments are deduced from manual analyses of dialogue corpora. For example, the two following word sequences “about four o’clock” and “at half past two in the morning” are both an instance of the conceptual segment named hour.
In a train schedule inquiry system, the user’s utterance “I would like a train to Paris tomorrow at nine” can be segmented into four conceptual segments as follow:

“I would like a train to Paris err well err at nine”

request destination filler hour

In our model, the phrase “I would like a train” is identified as an illocutionary act of request although its surface form represents a modality according to the speech act theory.

The language model is represented by a graph of conceptual segments modeled by a Hidden Markov Model (HMM) where states are the conceptual segments. Each conceptual segment is also represented by a HMM where the states represent word classes. Word classes were introduced to reduce the model complexity and the number of parameters to be estimated. A word class can be an emission state in different conceptual segments.

A simplified conceptual segment of destination is shown in the figure 1: all the destination prepositions are merged into the same word class (C1), and all the city names in the town word class (C2). Each state is associated with the set of lexical units which can be emitted (word class lexicon). The union of all word class lexicons defines the understanding lexicon.

<table>
<thead>
<tr>
<th>I</th>
<th>C1</th>
<th>C2</th>
<th>Toulouse</th>
<th>Paris</th>
</tr>
</thead>
<tbody>
<tr>
<td>toward</td>
<td>to...</td>
<td></td>
<td>Toulouse</td>
<td>Paris...</td>
</tr>
</tbody>
</table>

Figure 1: Example of conceptual segment.

2.2. Objectives

In order to enable the understanding system to detect and to interpret misrecognitions, our method consists in adapting the basic stochastic and conceptual language model. Firstly, we define our notion of OOV words at the speech understanding level: a word is considered as OOV for a word class when it is not included in the word class lexicon.

For instance, the user’s utterance “Je vais à Tours” (I go to Tours) could be recognized as “Je vais à jour” (I go to day). According to the notion of OOV words, the word ‘jour’ is known for the language model but it is considered as an OOV word for the lexicon of the town word class. In this case, it would be interesting to be able to understand that the user is speaking about a town.

This clue on the nature of awaited information would allow to inform the user about what has not been understood and thus to manage better the continuation of the dialogue. To the contrary, if the misrecognized word is in fact useless for the understanding, its presence must not disrupt the interpretation of the rest of the utterance.

2.3. Modeling

The modeling principle proposed in this paper consists in adding to each word class (including the word classes of the filler conceptual segment) the OOV word with its emission probability. These emission probabilities could be automatically trained from speech recognition output or manually defined by a speech expert. The OOV word is modeled in the same way than any other word of the word class.

With this notion of OOV words, our model is able to handle both words labeled OOV by the speech recognition and misrecognitions. In this case, the difficulty is that a misrecognized word belongs to one or several wrong class(es): it is more difficult to detect and to interpret misrecognitions because of ambiguities due to the confusion of the pronounced word with another one. The context provided by the preceding and following words can allow the understanding process to raise the ambiguity.

Contrary to [4], the emission probability of an OOV word concerned all word classes and not only one semantic category. However, it is obvious that some word classes are most sensitive than others to emit an OOV word. In our model, this consists in increasing the emission probability of OOV words for these classes. Contrary to [3], [4], the OOV word label is not useful. Our model is able to handle in the same way misrecognitions and OOV words labeled by the recognition system.

In order to show the effect of our approach to interpret unknown or misrecognized words, we consider three models:

- **Baseline:** this one uses the model described in 2.1. In that case, the understanding process is unable to interpret an utterance containing an unknown word and a misrecognized word will be interpreted as it is.
- **Weighted free model:** all of the word classes have the same probability to emit OOV words.
- **Weighted model:** a subtle analysis of the corpus allowed to determine that the most part of the unknown (hence misrecognized) words are towns (about 54% of instances) or words regarded as useless for the understanding (38% of instances). This observation led us to define a model which probability of emitting OOV words is higher for the word classes of conceptual segment filler and for the class town.

3. APPLICATIVE FRAMEWORK

The application domain concerns a train schedule inquiry system. The DEMON platform [7], a real time spontaneous speech dialogue system in the ARISE LE3-4229 project (Automatic Railway Inquiry Systems for Europe) [8] was developed with the speech tools of PHILIPS for French language. The DEMON system gives over the phone actual timetable information for the connection of 600 French train stations.

The language model consists in 29 conceptual segments (including the filler conceptual segment) and in 100 word classes. The filler is represented by one emission state (this class is also named filler). The language model was trained on 4268 orthographic transcriptions of utterances manually labeled according to the conceptual model of the application. For the two others models, the emission probability has been defined in a subjective way, after analysis of the corpus.

We have chosen to represent the meaning utterance in the form of key-value pair set.

The understanding error rate is computed with the Levenshtein algorithm, from the comparison of the key-value pair reference elaborated by an expert and the key-value output
produced by the CACAO system. This rate is calculated by adding up the number of insertions, deletions and substitutions of key-value pairs. We distinguish substitutions where the attribute is different and substitutions where the attribute is exact but the value is wrong.

4. EVALUATION

The goal of this evaluation is to show the influence of model type. The corpus test contains 2542 utterances. The understanding input is the N-best solutions computed from the recognition word lattice. We have realized this test with four values for N (N = 1, 2, 3 or 5). For these four instances, the figure 2 shows the understanding error rates obtained for the three model implementations proposed. The number of recognition errors is significant in this corpus. If we consider only the best solution given by the recognition module, the word error rate is 40.5% and there are about 23% of substitutions.

Figure 2: Understanding error rate depending on the recognition number of solutions and the model type.

We see that the least good results are those obtained with the baseline, and that the best are those gained from the weighted model (figure 2). We also notice that the higher the number of solutions given by the recognition is, the weakest the difference between the three models is. This phenomenon is logic: the more numerous solutions are considered, the more likely there is belong those the one matching with the pronounced utterance. But one can judge by the contribution of misrecognized word processing only if there are recognition errors. On the other hand, this shows that our approach does not degrade the results obtained from the baseline.

5. TESTS ON MISRECOGNIZED TOWNS

The main goal of the tests described here is to evaluate the contribution of our model for the processing of unknown or misrecognized towns or train-stations.

5.1. Word ambiguity rate

Some misrecognized words are more difficult to detect and interpret than others. This difficulty is due to the degree of confusion with the actually understood word, called word ambiguity rate: the rate $\Theta(M)$ of the word $M$ express the difficulty to correctly interpret $M$ when $M$ is due to a recognition error.

The ambiguity rate of a word $M$ is closely tied with the language model. It depends on several factors:

- The number $N_C$ of word classes containing the word $M$ in their lexicon;
- The number $N_{CE}$ of conceptual segments having a word class that could emit the word $M$;
- The emission probability $E(c_i, M)$ of the word $M$ for all word classes $c_i$ where it is not regarded as OOV;
- For each conceptual segment $CS_i$ containing a word class where $M$ can be emitted, the maximum transition probabilities $T_{CS}(i, c_i)$ to go from initial state $i$ of the conceptual segment $CS_i$ to the class $c_i$ where the word $M$ can be emitted;
- For each conceptual segment $CS_i$ containing a word class that emits $M$, the transition probabilities $T_{CS}(i, c_i)$ of every conceptual segment $CS_i$ of the language model to this segment.

The ambiguity rate $\Theta(M)$ of word $M$ is given by the following formula: (1)

$$\Theta(M) = \sum_{c_i \in C_j} [E(c_i, M) T_{CS}(i, c_i) \max_{c_i} (T_{CS}(c_i, CS_j))]$$  (1)

The next section shows the influence of word ambiguity rate on the understanding performance.

5.2. Adaptation of test corpora

Our work hypothesis is that the semantic category of misrecognitions generates linguistic ambiguities. The aim is to study the influence of the ambiguity level according to the kind and the value of the emission probability of OOV words in specific word classes.

To reach this goal, we select all the utterances both correctly understood and related to the city name class (corresponding to the departure and destination conceptual segments) during the reference test. We choose five words to replace the town name and then simulate misrecognized word:

- Word 1 ($\Theta = 0$): the chosen word is not available in the understanding lexicon and it is a priori the least ambiguous. This word may be considered as the OOV label produced by the speech recognition process like in “[1], [2].
- Word 2 ($\Theta = 0.87%$): it belongs to one of the word classes of the date conceptual segment. It is the word ‘jour’ (day).
- Word 3 ($\Theta = 1.31%$): it is a number and it can be found in the number word class (one of emission states of date and hour conceptual segments).
- Word 4 ($\Theta = 1.59%$): this word belongs both to the filler word class and to one of the word classes of the response conceptual segment. It is the word ‘bon’, in English it may be ‘well’ or ‘OK’ in the response segment like in “oui, c’est bon” (“Yes, it is OK”).
• Word 5 ($\theta = 5.96\%$): it belongs to the filler word class but not to any other word classes. So, it can be confused with a non-significant word for the understanding. We chose the word ‘alors’ (then) which is very frequent in the corpus and then very difficult to interpret.

The ambiguity rate of these words strongly differs. The test was made five times with each of these five words. 1477 utterances constitute the corpus test. 1089 of these utterances contain only one OOV word. These utterances correspond to 6332 conceptual segments (i.e. around four conceptual segments per utterance) and to 4035 key-value pairs.

5.3. Results and discussion

The figure 3 displays the results obtained depending on the word ambiguity rate and the model chosen.

![Figure 3: Tests on misrecognized towns.](image)

The results show that the weighted free model influence is not significant for misrecognized words, but very high for the towns unknown to the application (for the word 1). Remind that, with the baseline, it is impossible to interpret an utterance containing an unknown word, which explain the 100% error rate for the first word. In the same way, a misrecognized word will not be interpreted accurately. However, despite the presence of a misrecognized word, the rest of the utterance can be correctly understood. We notice that, whatever the ambiguity rate of the mingled word is, the understanding error rates are equivalent for the two first models (about 50% for the last four words). Consequently, the weighted free model does not improve the performances for misrecognized towns. On the other hand, the third model appreciably improves the results, whether the town is unknown or misrecognized.

Concerning the weighted model, the performance strongly depends on the ambiguity rate of the misrecognized word: the understanding error rate increase from 0.8% for the word 1 (null ambiguity rate) to 33.7% for the word 5 (highest ambiguity rate).

From a detailed analysis of the error’s type, we have noticed that, on the whole, the interpretation mistakes don’t have repercussion on the rest of utterance. The majority of these errors concern the deletions of key-value pairs which are related to the towns. Some key-value pairs are sometimes erased: they are “absorbed” by the filler conceptual segment. The third kind of error occurs when the understanding process could not detect it wasn’t the correct word, which generates confusions with some others key-value pairs.

6. CONCLUSIONS

We have proposed an enlargement to the conceptual stochastic model for the treatment of misrecognized words. This one also permits to interpret out-of-vocabulary labeled words. We have seen that the processing of misrecognized words is quite critical and we have noticed that some words are far easier to identify as being misrecognized than some others. This led us to define an ambiguity rate for words. The results obtained depend directly on the ambiguity rate of misrecognized words: indeed, the higher this rate is, the harder it is to interpret accurately the word and the conceptual segment containing it. However, this model has major limit: it is impossible to detect that a word is misrecognized if it belongs to the same class word that the pronounced word (for example when a town is merged with another one). A method to remove this difficulty would be to take into account the confidence scores given by the recognition module on the words in order to detect that the word is misrecognized.

7. ACKNOWLEDGEMENTS

The authors acknowledge Martine de Calmès to have made available training and testing corpora.

8. REFERENCES