CORRECTING ASR OUTPUTS: SPECIFIC SOLUTIONS TO SPECIFIC ERRORS IN FRENCH

Richard Dufour, Yannick Estève
LIUM, Université du Maine
Le Mans, France
firstname.lastname@lium.univ-lemans.fr

ABSTRACT

Automatic speech recognition (ASR) systems are used in a large number of applications, in spite of the inevitable recognition errors. We propose in this study a pragmatic approach to automatically repair ASR outputs taking into account linguistic and acoustic information, using formal rules or stochastic methods. The proposed strategy consists in developing a specific correcting solution for each specific kind of errors. In this paper, we apply this strategy on two case studies specific to French language. We show that it is possible, on automatic transcriptions of French broadcast news, to decrease the error rate of a specific error by 11.4% (relative) in one of two the case studies, and 86.4% in the other one. These results are encouraging and show the interest of developing more specific solutions to cover a larger part of errors in a future work.

Index Terms— Automatic speech recognition, error correction, homophones, language modelling

1. INTRODUCTION

Automatic speech recognition (ASR) systems are increasingly efficient. Their actual performances are sufficient to be used in a large number of applications (speech human-machine dialog, indexing, information retrieval, ...).

But ASR errors are inevitable. Errors changing the sense of a sentence are very annoying for the most part of applications using ASR, because they prohibit correct feedbacks from these applications. Other kinds of errors, which do not prevent the speech understanding, are often neglected because they are not critical for the correct operation of such applications. These errors, for example in English, can be errors of agreement in number.

For some other applications, as subtitling for deaf people or assisted manual transcription [1], these errors have a bigger importance: for the first one, repetition of errors, even if they don’t modify the sense of a sentence, are very tiring for the final user; for the second one, which consists in producing an entirely correct transcription, these errors reduce the gain of productivity provided by the use of ASR system outputs.

As some languages, French contains a lot of homophonous words, particularly for different inflected forms of a same word. This kind of ASR errors is frequent. So, in this context, it should be very interesting to have a post-processing method to correct such errors.

In the literature, we can find propositions to repair ASR errors or make the applications robust on these errors [2, 3]. Usually, propositions to repair ASR errors tend to be generals and try to repair every kinds of errors [4]. They also can take into consideration some particularities of the focused application, for example the dialog history [4].

In this paper, we propose a different approach consisting in building a specific correcting solution for each specific error. Such kind of approach was yet proposed in [5] for a use inside the ASR decoder to handle different language models. Here, we propose to use it in order to repair some errors in a post-processing level.

This approach is very pragmatic: it consists to manually analyze the most frequent errors, particularly the most frequent confusion pairs. These errors on different words can be gather into the same specific kind of errors (for example, as we will see below, the errors due to homophonous inflected forms of past participles in French), or can be processed confusion pair by confusion pair. In the different solutions proposed for the different kinds of errors, heterogeneous tools can be used, as formal rules or stochastic methods, which can be built from heterogeneous data, as linguistic knowledge, but also acoustic information provided by the ASR system. The use of acoustic information to repair some ASR errors in a post-processing level is another contribution of this paper.

This paper focuses on French language. Two case studies are presented, consisting in correcting specific errors due to two particularities of French. The next section presents some specificities of French. Then the proposed approach is detailed in section 3, before the presentation of the tools used during this work. Last, experiments are described and results are presented.

This research was supported by the ANR (Agence Nationale de la Recherche) under contract number ANR-06-MDCA-006.
2. SOME SPECIFICITIES OF FRENCH LANGUAGE

The correspondence of gender, number (and/or person) is one of the most difficult aspects of the French language. French is an inflected language. A great difficulty for ASR systems (and for some people) is that in a lot of cases, the different inflected forms of a word are homophones and only language models can distinguish them.

Agreements in gender or number are not always well-modeled by the n-gram language model used in ASR system because the length of the constraints modeled by n-gram model can be not sufficient, but also because of some weakness in the language model due to the lack of data or the vocabulary size: [6] shows that the number of distinct words in French in a vocabulary must be the double than English to obtain the same word coverage.

Moreover, some complex grammatical rules can be hardly modeled with a n-gram language model.

3. PROPOSED APPROACH

In this paper, we focus on the errors due to the homophones inflected forms of past participles, and we focus on the errors concerning the words 'vingt/vingts' (twenty) and 'cent/cents' (hundred) both. These errors belong to the most frequent errors produced by our ASR system, according to the analysis of the confusion pairs computed by the NIST SCLITE software.

To repair errors, we seek to use formal rules whenever it could be possible. But this approach cannot be the only one. In particular, formal rules are not very robust with the errors existing in the lexical context of a targeted word. Thus, when it is possible to establish a formal rule, we do. If not, we try to use a statistical method, for example a statistical classifier which is built from a training corpus, in order to make a decision for calling into question a hypothesis word proposed by an ASR system and to propose a better solution.

Therefore, we do not try to resolve every error, but we target errors that appear correctable. To develop a specific solution, different knowledge bases can be used. Lexical information, of course, but acoustic information, part-of-speech (POS), or other information levels can be useful. So, information coming from the ASR can be used, but also POS tagger or other available tools or knowledge.

POS will be very useful to build formal rules, and give the classifier the possibility to get richer information to detect and correct agreement errors. Acoustic information can be the pronunciation variant used by the ASR system to recognize a word. It is possible to link these pronunciations to other linguistic information (e.g. gender and number) to guide choices when correcting them.

Taking the example of a correction proposed by a statistical tool about the masculine/singular word form “transcrit” associated to phonemes [t r an s c r i] in French. Assuming that the word is considered false by our statistical system (classifier), it proposes feminine / singular form. This would give “transcrite” [t r an s c r i t]. But if the acoustic information is not consistent, the proposition is not retained. Moreover, acoustic information allows sometimes to find the acoustically discriminant word which brings the information on the gender or the number concerning the targeted word in the utterance. According to information made available to train the classifier, this will select the most likely class. Classes are determine in function of the targeted error to repair. For example, if we want to know the number (singular or plural) of a word: singular will be considered as a class and plural as another one. The classifier will test the word and propose a number (singular or plural) for this word. When the classifier and the ASR system disagree, we replace the ASR hypothesis word by its inflected form corresponding to the number given by the classifier.

The approach is pragmatic and is divided in two fronts, on the one hand the establishment of formal rules, on the other hand the use of statistical methods (classification) on specific problems. The method giving the best results on development data is conserved.

4. TOOLS

Experiments on speech recognition were carried out by using the LIUM ASR system, based on the CMU Sphinx 3.x decoder, described in [7]. It is a three-pass system: the first pass uses a trigram language model and generic acoustic models (one for each of the four gender/band conditions — female/male + studio/telephone), the second pass uses the best hypothesis of the first pass to adapt the acoustic models using SAT and CMLLR, and the last pass consists in rescoring with a quadrigram language model a word-graph generated during the second pass. This system ranked second in the French ASR evaluation campaign ESTER, and was the best open source system. On the test data, the system reaches a 22.2% in terms of word error rate (WER). Notice that the substitution error rate is 13.6%: this is important because the approach proposed in this paper try to repair only some substitution errors (insertions and deletions are not targeted at this time).

POS are associated to hypothesis words using the lia_tagg tagger [8] distributed under GPL license.

We choose to use the state-of-art classifier BoosTexter [9] which allows to build a powerful classification from text and continuous values.

5. EXPERIMENTAL DATA

Data used for the experiment comes from the ESTER evaluation campaign [10] on rich transcription of French broadcast news. This corresponds to 100 hours of manually transcribed recordings.
The ASR system used for the experiment was developed through the transcribed data provided by the ESTER campaign. Data are divided into three corpora: train data, development data, and test data.

The statistical classifier was built from the manual transcriptions of the train data. Notice that the language models used in the ASR were, in a minor part, estimated from the same train data (and with newspaper articles from the "Le Monde" newspaper, not used to train the classifier).

Table 1 presents the sizes of the three corpora.

<table>
<thead>
<tr>
<th></th>
<th>Training</th>
<th>Development</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words</td>
<td>840K</td>
<td>87K</td>
<td>114K</td>
</tr>
<tr>
<td>Time</td>
<td>73 hours</td>
<td>7 hours</td>
<td>10 hours</td>
</tr>
</tbody>
</table>

Table 1. Size (words and time) of the three corpora

Acoustic information (phonemes associated with words) are in the pronunciation dictionary of the ASR decoder. This dictionary contains about 165K pronunciation variants for 62K words.

To enrich the information about words proposed by the ASR system, all the grammatical forms of a word are need. This information was get from lexique3 database, which contains 135K French words. It gives representation and phonemic agreement, syllabation, gender, number, and word lemma associated.

6. EXPERIMENTS

We focus our work on localized and specific agreement errors to the French language. We choose frequent errors coming from our ASR system. Thus, as a first step, we look through language rules whether it is possible to remove agreement errors in number for words cent (singular cent / plural cents) and vingt (singular vingt / plural vingts), which use the same rule (bad modeled by ASR n-gram models).

In a second step, we focus on agreement errors of French past participles which can be broken up to four different forms (depending on gender and number). The difficulty lies in the fact that these forms can be phonetically identical, and it is extremely difficult to apply linguistic rule to get the correct spelling. We first tag the words contained in the left- and right- contexts of the targeted past participle using the POS tagger. This information was used during the training phase of the classifier in order to general the observed events. Moreover, it contains some morpho-syntactic information.

If the gender or the number provided by the classifier are different than the ASR proposition, the classifier proposition is retained only if the pronunciation of the new inflected form is the same than the one proposed by the ASR.

7. RESULTS

To find out if our approach helped to repair the focused errors, we modify directly transcriptions generated by ASR in order to compare the error rates. The first error rate was computed on the development data, which helped to calibrate our system, and the second rate was computed on the test data, to assess its effectiveness. We present in table 2 the rates of agreement errors on words “cent” and “vingt” before (Baseline) and after correction the use of the linguistic formal rule (Correction).

Table 2. Rate of agreement errors on words “cent” and “vingt”.

<table>
<thead>
<tr>
<th></th>
<th>Development</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>16.7%</td>
<td>6.6%</td>
</tr>
<tr>
<td>Correction</td>
<td>0.6%</td>
<td>0.9%</td>
</tr>
<tr>
<td>Relative gain</td>
<td>-96.7%</td>
<td>-86.4%</td>
</tr>
</tbody>
</table>

The impact of the formal rule is positive. During the development phase, the relative gain was really high, with a correction of 87 words on a total of 90. So we test our system on test data and observe that only six words are still badly recognized after correction (due to lack of historic information or wrong word choosen during ASR decoding) but do not addition of errors. We managed to correct 28 errors of “cent” and 13 of “vingt”. Although the relative gain is less important than development phase, it still remains interesting. By using a linguistic rule, in post-processing of the ASR outputs, we can see that it is possible to correct agreement errors, with a relative gain of 86.4% on those words.

We now focus on the results of the second part of the experiment: the correction of homophonous inflected forms of past participles using a statistical method. The table 3 compares the rate of past participles that have an agreement error before (Baseline) and after the use of the statistical method (Correction), in order to obtain the relative gain.

Table 3. Rate of agreement errors due to past participles.

<table>
<thead>
<tr>
<th></th>
<th>Development</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>10.1%</td>
<td>12.9%</td>
</tr>
<tr>
<td>Correction</td>
<td>8.1%</td>
<td>11.4%</td>
</tr>
<tr>
<td>Relative gain</td>
<td>-19.6%</td>
<td>-11.4%</td>
</tr>
</tbody>
</table>

Given the positive results that we had on development data (58 words corrected on a total of 296), we decided to extend our system to the test data. During the test phase, the classifier detected 253 past participles containing agreement errors. Thanks to the inclusion of acoustic parameters, 80 (31.62%) past participles which had to be modified were ignored. This acoustic constraint permits to improve the relative
gain by 2.6% absolute.

The figure 1 summarizes the performances of the statistical method coupled with the acoustic constraint on the past participles which it has modified. There are reparations (well modified), introduction of new errors (false alarm), and finally modification without effect (these past participles cannot be repaired because the correct words are not one of their inflected forms).

We note that the proportion of past participles well adjusted (70 words) is far greater than wrong past participles introduced (27 words). Finally, this method allows to correct 43 past participles due to an agreement error (relative gain of 11.41%).

Those methods (linguistic rules and statistical method, combined with acoustic parameters) allows to correct agreement errors on transcriptions from ASR without modifying the decoder. This preliminary work focusing on a limited number of rules shows the possibility of correcting this kind of errors. Moreover, to apply those methods, we only need 6 minutes 16” to be achieved, which is negligible compared to the amount of data processed.

8. CONCLUSION

In this paper, we have proposed a strategy consisting on building specific solutions to repair specific ASR errors. We have applied this strategy through two case studies, which are two of the most frequent errors produced by our ASR system on French broadcast news recordings.

First, we have proposed a solution to repair some errors due to a particular rule concerning the agreement in number of the homophones words 'vingt/vingts' (twenty) and 'cent/cent's' (hundred). This solution uses formal rules and allows to reduce the error rate for this kind of errors by 86.4% relative on the ESTER test corpus.

Secondly, we have proposed a solution to repair some errors due to the homophones inflected forms of past participles. In this case, a stochastic method using the boostexter classifier, allows to reduce the error rate for this kind of errors by 11% relative.

Of course, these two improvements are not sufficient to reduce significantly the global word error rate. But we have showed that it is possible to decrease drastically the error rate on specific errors. In the future, we will develop such specific and pragmatic correcting solutions in order to cover a larger part of errors.

Last, the solutions developed to repair errors does not seem dependent on ASR systems. In the future, we will make experiments to evaluate the reusability of these solutions.

9. REFERENCES