AUTOMATIC ERROR REGION DETECTION AND CHARACTERIZATION IN LVCSR TRANSCRIPTIONS OF TV NEWS SHOWS

Richard Dufour, Géraldine Damnati, Delphine Charlet

Orange Labs, France Telecom, Lannion (France)
firstname.lastname@orange.com

ABSTRACT

This paper addresses the issue of error region detection and characterization in LVCSR transcriptions. It is a well-known phenomenon that errors are not independent and tend to co-occur in automatic transcriptions. We are interested in automatically detecting these so-called error regions. Additionally, in the context of information extraction in TVBN shows, being able to automatically characterize detected error regions is a crucial step towards the definition of suitable recovery strategies. In this paper we propose to classify error regions in four classes with a particular focus on errors on person names. We propose several sequential detection + classification approaches and an integrated sequence labeling approach. We show that our best classification system can reach 70% classification accuracy on automatically detected error regions. Additionally, the overall system is able to detect and correctly characterize 29.6% of error region corresponding to a person name with a precision of 61.9%.

Index Terms— Error region detection, Error characterization, Automatic classification, LVCSR.

1. INTRODUCTION

In the context of large vocabulary continuous speech recognition (LVCSR), automatic speech recognition (ASR) systems can now provide transcriptions with good performance, allowing them to be integrated into many applications. However, some of the remaining errors can still be a problem for certain applicative domains, such as information extraction. In this article, the purpose is to identify and characterize errors inside automatic transcriptions. Not only are we interested in detecting erroneous words but the purpose is to detect and characterize error regions (i.e. clusters of consecutive errors).

Traditionally, error detection has been handled through the definition of confidence measures (CM) representing the probability of a word to be correct [1]. Applying a threshold on this score allows the system to be tuned at a given operating point which can be suitably chosen depending on the applicative context (high recall or precision). CM can be seen as a binary classifier that classifies words into correct/incorrect but they are usually evaluated in their ability to determine if a word is correct. However, for tasks where word error rates are low, the binary classification task is typically an unbalanced data binary classification task and evaluation focused on the majority class (correct words) tends to hide the ability of the classifier to handle the minority class (erroneous words). In this paper we are interested in detecting errors in the context of TV Broadcast News (TVBN) shows with unbalanced repartition of input data (in favor of correct words). The evaluation issue will be addressed and we will focus on the ability of our system to effectively detect errors.

It has been already observed in the literature that ASR errors are not independently distributed, nor across the data (some parts of the shows, some speakers generate more errors than others), nor with respect to other errors. Some phenomena generate several consecutive errors (so-called *error regions* in this paper). This can happen for many reasons, such as an unknown long word substituted by several short words, a substitution that propagates to adjacent words because of the language model, or still bad acoustic conditions for which the decoder provides a completely erroneous output. In [2], authors analyzed Broadcast News (BN) and Conversation Telephone Speech (CTS) transcription errors and concluded that two-third of the errors appear in a group ($n \geq 2$ consecutive errors). In this paper we are interested in detecting and characterizing those *error regions* in order to be able to define suitable recovery strategies. The addressed problem is then a sequence labeling problem of ASR outputs.

Beyond error detection we are interested in automatically characterizing error causes. In fact, not all error causes have the same impact in a given applicative context. This automatic characterization of an error would allow to decide whether to ignore it or eventually to define suitable strategies to correct it. From the analysis point of view, several studies have provided detailed posterior analysis of error causes. [2] highlighted that the majority of errors in BN transcriptions in English are related to Named Entities. In [3], authors highlighted that homophones in French language are very frequent and represent an important source of errors in ASR outputs. Acoustic conditions or other linguistic phenomena such as disfluencies are also usually analyzed as error causes. From the automatic characterization point-of-view however, many studies focused on detecting and correcting Out-Of-Vocabulary (OOV) errors, whose behavior and impact differ from other errors [4]. Specific strategies have been proposed to detect OOV words by using for example an hybrid word and sub-word language model [5, 6] or a semantic class language model [7]. In [8], authors focus on error regions generated by OOV words and propose a method which takes into account neighboring contextual information instead of only considering the local region of OOV errors. Their experimental corpus has been designed by only keeping meaningful OOVs and excluding the ones with less than 4 phones and region boundaries are supposed to be known in advance by aligning confusion networks with the reference transcription. The *OOV / IV* classification task for these error regions yields a missed OOV rate of 28.4% at 10% false alarm rate. In highly inflected languages, OOV words can be of various nature. Not only proper nouns involved in Named Entities are OOV sources but also inflections of a given lemma or rare words in the general case. Thus we have chosen not to focus on OOV as a class but to define more relevant classes for our applicative context, each of them potentially containing OOV words. For example, correctly recognizing person names is essential, no matter if the error is due to an OOV word, the result is that the person name was not correctly transcribed.

Experimental protocol and focused classes are described in section 2. We propose to study in section 3 the distribution of error sequences inside TV show transcriptions. We then propose several methods to both detect and characterize error regions, and experi-

ments will evaluate them in the last section depending on the detection, the classification and the overall sequence labeling task.

2. EXPERIMENTAL DATA AND PROTOCOL

2.1. Database description

A corpus of 24 French TVBN shows recorded from October 2008 to January 2009 has been manually transcribed. Person names have also been manually annotated. The total duration of this corpus is around 8 hours of speech for around 84k words. The corpus has been split into a training corpus (15 shows/5 hours) and a test corpus (9 shows/3 hours). It includes TVBN shows from 7 French generalist channels and contains planned studio speech as well as interviews, live reports. As we process complete shows, Broadcast Conversation (BC) portions are kept in the experiments. ASR is performed using the VoxSigma speech recognizer v3.5 from Vocapia Research, based on LIMSI technology [9]. The dictionary contains 65k words, including 22k proper nouns. The word error rate (WER) on the overall corpus is 15.9%. The 13,373 transcription errors decompose in 6,731 substitutions, 4,299 deletions and 2,343 insertions. Detecting deletion errors is a difficult task and we chose to focus in this work on substitutions and insertions. Thus our material is composed of 82,352 hypothesized words, containing 9,074 erroneous words. In this context detecting errors is typically a minority class detection problem. We use the word confidence measure provided by the ASR system: posterior probabilities from word lattices. It is a performant CM from the classical CM evaluation point of view as it results in a 0.36 Normalized Cross Entropy score over the whole corpus.

2.2. Definition of error classes

We chose to focus on 4 types of error sources that are determined from the alignment between automatic and manual transcriptions. The alignment is performed with the NIST Sclite tool[1], usually used to evaluate automatic transcriptions. Firstly, the *Person name (PN)* class is particularly studied, as this information is essential in many applications of information extraction. We are also interested in *Other proper noun (OPN)* class, which still contains very useful information. The *Homophone (H)* class is studied in order to retrieve a frequent phenomenon of French language. Finally, all the remaining errors are clustered in the *Other (O)* class.

In order to determine the *Homophone (H)* errors, an additional lexicon has been used in order to compare phonetic representation of reference and hypothesized words: identical phonemes indicate here the presence of homophones.

2.3. Deriving error region reference labels

Throughout this work, an error region is defined as a sequence of consecutive errors. In order to determine the reference label of a given error region, the sequence of erroneous words is compared to the corresponding sequence of reference words. Even if deletion errors are not considered in the detection task, they are taken into consideration for the reference labeling of error regions. In fact, when two reference words are recognized as a single word which corresponds to one of the two reference words, the Sclite alignment algorithm will determine that the first word is deleted and the second word is substituted. In order to decide for the substitution error category, we need to look at the properties of the two reference words. This is the reason why, when calculating the corresponding sequence of reference words, we also include adjacent omitted words. Finally a sequence of n_h erroneously hypothesized words is aligned with a sequence of n_r reference words where n_h and n_r are not necessarily equal. To determine the error region label, an order of importance is associated for each class. The priority is assigned as follows and an error region corresponds to the class:

1. *Person name (PN)* if one of the reference words is part of a person name.
2. *Other proper noun (OPN)* if no word in the reference sequence is part of a proper name and if one word of the reference sequence is a proper noun.
3. *Homophone (H)* if no word in the reference sequence is a proper noun and one of the erroneous words of the region has been aligned with an homophone word in the reference.
4. *Other (O)* otherwise.

If one type of error is found in the error region, the entire error region takes the error class label depending on its priority (for example, if the error region contains three errors, including one *person name* error, and two *others*, the region has the *person name* label).

3. ERROR REGIONS INSIDE LVCSR TRANSCRIPTS: A QUANTITATIVE STUDY

To assess the interest of recognizing and categorizing error regions, we present in this section a quantitative study which will analyze error regions depending on various parameters. Firstly, figure 1 shows repartition of transcription errors depending on length of sequence errors. Two sources of data are presented: the error repartition per word and per error region (for example, two consecutive errors count for one region). We can see that more than 25% of errors are isolated errors and 55% of regions are singletons (histogram 1). These results, made by analyzing TVBN transcriptions, are close to the ones presented by [10] which found that 30% of errors occurred in isolation for dictation systems. On the overall, an average of 1.7 consecutive errors has been observed. When focusing on multiple error regions, representing 75% of the misrecognized words, the average length of these regions is 2.2 words. This score comforts the fact that it is interesting to try to detect these regions.

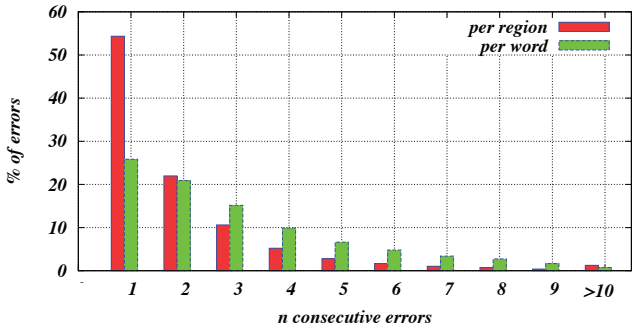

Fig. 1. *Repartition (per word and per region) of transcription errors depending on length of sequence errors.*

Table 1. *Number of error regions, average number of consecutive errors and OOV region proportion depending on the 4 error classes.*

	PN	*OPN*	*H*	*O*	*All*
Error regions	359	301	1,489	3,024	5,173
Error region length	2.0	2.1	1.5	1.7	1.7
% OOV	51.5%	31.9%	6.9%	6.6%	11.2%

Table 1 presents the repartition of error regions for the 4 error classes, the average number of consecutive errors, and the OOV proportion per region class. The high number of homophone error regions justify our choice to isolate this error for French language. In fact, by looking more closely to this error type, we found that 74% of these errors are due to inflected forms of a word (same lemma but different spelling depending on the context). *Person name* and *Other proper noun* error regions are longer on average (more than 55% of *PN* and *OPN* error regions are multiple error regions with $n \geq 2$).

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

At the opposite, more than 60% of *homophone* and *other* error regions are singletons. The last line shows that 48.5% of the *person name* error regions are not due to an OOV word, which reinforces our idea to detect error regions and not only OOV ones.

4. EXTRACTION AND CHARACTERIZATION OF REGION ERRORS

4.1. Sequential approach

As mentioned in the introduction, simultaneously detecting and characterizing error regions can be seen as a sequence labeling task. We firstly propose a sequential approach which, in a first step, consists in segmenting transcriptions into error region / correct region, and in a second step associating these error regions with a class.

4.1.1. Error region segmentation

We propose 3 different approaches for segmenting error regions. Firstly, we use the classical approach and propose to segment error regions by applying a threshold θb on *a posteriori* word confidence scores provided by ASR system. In fact, the consecutive words detected as a possible error ($< \theta b$) will be considered as an error region. This method will be called *Baseline*.

Applying a single threshold on CM may not be sufficient as consecutive errors are not necessarily all associated with a low confidence score. In order to relax the constraint, we introduce an automaton with two thresholds, as presented in figure 2. Each word of a sentence is analyzed: in the *Correct* state, the word is considered as correctly recognized, while the *Error* state detects the word as incorrect. The two thresholds on CM are very important, since the threshold θerr permits to change from *Correct* to *Error* state, or to stay in the *Correct* state, and conversely for the threshold θcor. The last constraint is that θerr should be inferior to θcor. The automaton will be used for each sentence at both side (from left to right and vice versa) to capture errors not found in a way (due to 2 thresholds). Higher order automata have been implemented but yielded too large regions. In this approach, we are not only focused on the current CM, but on the surrounding ones. For example, it is possible to stay in the *Error* state if the confidence score is between θerr and θcor. It would not have been possible with only one threshold, and thus should catch larger error regions.

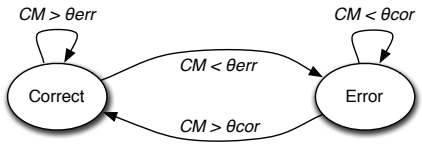

Fig. 2. *Capturing error regions using an automaton with two thresholds (θerr / θcor) and confidence measures (CM) of transcribed words.*

Finally, we propose to use Conditional Random Fields (CRF) [11], a statistical method allowing to segment and label sequence data. As presented in [12], CRF can be used to improve *a posteriori* confidence scores by adding extra features (words, POS tags...) and methods to re-estimate them. The advantage is that this method uses various sources of information about surrounding words to detect regions: bigram words, Part-Of-Speech (POS) tags and syntactic chunks [2], CM and the duration of current, previous and next words.

4.1.2. Error class labeling

After, finding these error regions, we propose to associate them with one of the 4 error classes described in section 2 using a classification method. We chose to use *Icsiboost*[3], a large-margin classifier based on the *AdaBoost* algorithm. Various features are used: region words (bigram), trigram POS tags and syntactic chunks, the number of words of the region, quadrigram on the 5 previous words, the duration and the average of CMs of the speaker turn and the average number of syllable per word.

4.2. Integrated approach

As CRF can segment and label sequence data, we propose to use that method to directly retrieve error regions and label them with one of the 4 focused error classes instead of detecting regions and then categorize them. The same features than the one described for the sequential CRF approach will be used.

Finally, the previously proposed approaches should provide different error regions and classes. So, we propose a last solution which consists in combining all these outputs in merging error regions with the boolean operator "OR", and after choosing the error region class still depending on the priority defined in section 2.

5. EXPERIMENTS

5.1. Error region detection task

Firstly, we are interested in evaluating our proposed methods on the task of detecting error regions. We propose to use the classical precision/recall metric. Indeed, we are not particularly interested in retrieving the exact error regions, but pointing out places in transcriptions where an error region could appear: if a detected region overlaps with true error region, the detection is considered as correct even if region frontiers are not exactly retrieved. Table 2 presents precision, recall and F-measure obtained for the error region detection task with our methods and average length of detected regions.

Table 2. *Recall, Precision and F-measure for the error region detection task and average length of detected regions.*

	Recall	*Precision*	*F-measure*	*Avg length*
Sequential Baseline	24.0	78.0	36.7	1.2
Sequential Autom.	27.3	**87.9**	41.6	1.7
Sequential CRF	43.1	82.1	**56.5**	2.2
Integrated CRF	37.3	78.9	50.6	2.2
Fusion	**48.2**	79.6	**60.1**	2.3

We can see that *Sequential CRF* is the best sequential method for this task, with a recall of 43.1% and a precision of 82.1%. *Sequential Baseline* is not sufficient to capture error regions as only 1.2 consecutive errors are captured on average. *Sequential CRF* method is also better than the integrated one, which only reaches a 50.6% F-measure score. This could be explained by the fact that the integrated method is trained on more error classes, while evaluated on the same binary task. Finally, the *Fusion* is also very interesting, with the best recall (48.2%) and F-measure (60.1%) of all the approaches: we observed that regions of $n \geq 2$ errors are better detected with the CRF methods, while others better detect single error regions. However, the recall rate remains under 50%: more than half of the error regions are not detected. To ensure that the flexible evaluation constraint does not privilege the fusion approach, we also evaluated error region detection with a strict constraint (only the precise frontiers are considered correct). The same tendency could be observed, the *Fusion* is still the most performant approach. However, with this strict constraint, best results reach a 21.7% recall and a 35.9% precision. Precise frontier detection is not yet reliable and detecting error regions remains a difficult task.

5.2. Error region classification task

In these second experiments, we focus on the evaluation of error region classification task. Classification is evaluated in terms of classification accuracy: we want to know if the detected error regions are correctly labeled with the correct error class. The missed error regions, i.e. which are present in the reference but have not not been

[2] Lia_tagg: http://pageperso.lif.univ-mrs.fr/∼frederic.bechet

[3] http://code.google.com/p/icsiboost

detected, are ignored. A detected error region can either be a good detection or a false alarm. Table 3 presents the number of detected regions, overall classification accuracy and correct classification rate of the 4 error classes.

Table 3. *Number of detected regions, overall classification accuracy (All) and correct classification rate of the 4 error classes.*

	Nb regions	*All*	*PN*	*OPN*	*H*	*O*
Seq. Baseline	324	69.7	20.5	0.0	22.8	**96.7**
Seq. Autom.	369	69.8	24.2	4.1	21.8	96.2
Seq. CRF	554	66.3	30.3	0.0	19.5	96.6
Integ. CRF	512	**70.8**	51.5	13.5	**66.3**	80.3
Fusion	652	69.7	**53.4**	**15.2**	56.2	81.6

We can see that the best overall classification accuracy is obtained with the *Integrated CRF* method which reaches 70.8%. This method is also very accurate to label *Homophones (H)* (66.3%). This is not very surprising since this method is designed to directly detect and label region errors. It is also interesting to note that *Sequential CRF*, which has the best F-measure for the error region detection task, has here the lowest classification rate in the labeling task. It seems that the detected regions retrieved by this method are more difficult to classify. We observed that results are very low on *OPNs* when focusing on sequential approaches. The *Icsiboost* method is not efficient for this minority class. Finally, the *Fusion* approach allows to improve the *PN* (53.4%) and *OPN* (15.2%) performance.

Considering FA (falsely detected error regions), we measured with the *fusion* approach that only 3% of the 334 FA are classified as PN and 0.6% as OPN. This is an interesting result meaning that the classification process is robust to FA and that falsely detected error regions are not harmful from an applicative point of view.

5.3. Overall process evaluation

Finally, the last evaluation seeks to evaluate the complete detection and classification task. Table 4 presents performance in terms of recall and precision obtained for the 4 error classes.

Table 4. *Recall/Precision of the proposed methods applied to the 4 error classes.*

	PN	*OPN*	*H*	*O*
Seq. Baseline	6.1/80.0	0.0/0.0	4.2/41.1	25.1/47.4
Seq. Autom.	7.6/76.9	1.0/100	4.8/45.6	28.5/58.8
Seq. CRF	**17.4/60.5**	0.0/0.0	7.6/46.1	42.0/52.7
Integ. CRF	**25.8/59.7**	5.1/26.3	21.8/43.7	30.4/65.1
Fusion	**29.6/61.9**	7.1/36.8	24.4/42.6	40.6/61

By focusing on the sequential approaches, we can see that the performance is globally low, except for the *Other* class. This is not really surprising since this is the most represented error class. We can also note that the *Integrated CRF* method outperforms any of the sequential ones. Finally, with the *Fusion* approach, an improvement is observed, particularly on proper nouns (PN and OPN).

The overall detection and characterization process allows to detect 29.4% of PN with a precision of 61.9%. It is particularly interesting since we especially want to detect person names. Furthermore, 70% of person name occurrences in the corpus are uttered by the *anchor* speaker (main journalist presenting the show) while anchor speaker turns represent 25% of the total amount of speaker turns and 30% of uttered words (see [13] for a detailed analysis of the corpus in terms of speaker roles). This is the reason why we are particularly interested in evaluating our system performance for the subset of anchor speaker turns. This subset gathers 63% of the total error regions made on the PN. We obtained encouraging results, since we achieved 37.4% in recall and 79.5% in precision, meaning that we can detect 37.4% of person names uttered by the *anchor* speaker that were misrecognized.

6. CONCLUSION AND PERSPECTIVES

In this paper, we proposed to study LVCSR transcription errors in TVBN. We chose to address the twofold issue of error region detection and characterization. It is a well-known phenomenon that errors are not independent and tend to co-occur in automatic transcriptions. We have presented a quantitative study that illustrated this need for seeing errors as part of a region. In fact, only 25% of errors are isolated. Then, we proposed to classify error regions in 4 error classes: *Person name (PN)*, *Other proper noun (OPN)*, *Homophone (H)* and *Other (O)*. We particularly focused on the person names, since this class is essential for further information extraction processing. Several approaches have been proposed: sequential detection + classification process, integrated sequence labeling method. Our best classification system can reach a 70% classification accuracy on automatically detected error regions. Additionally, the overall system is able to detect and correctly characterize 29.6% of error regions corresponding to a person name with a precision of 61.9%. When focusing on the *anchor* speaker turns, the proposed method retrieves 37.4% of PN error regions with a 79.5% in precision. In a future work, we will explore new features, such as language model estimations, often used in OOV detection studies. Moreover, effort should be made on OPN detection class, since results are not yet satisfactory. The problem of the high number of missed region detection should be investigated. Finally, it could be interesting to analyze all error classes depending on various speaker roles.

7. REFERENCES

[1] H. Jiang, "Confidence measures for speech recognition: A survey," in *Speech Communication*, 2005, vol. 45, pp. 455–470.

[2] N. Duta, R. Schwartz, and J. Makhoul, "Analysis of the Errors Produced by the 2004 BBN Speech Recognition System in the DARPA EARS Evaluations," in *IEEE TASLP*, 2006, vol. 14, pp. 1745–1753.

[3] I. Vasilescu, M. Adda-Decker, L. Lamel, and P. Halle, "A perceptual investigation of speech transcription errors involving frequent near-homophones in French and American English," in *Interspeech*, Brighton, Angleterre, UK, 2009, pp. 144–147.

[4] P.C. Woodland, S.E. Johnson, P. Jourlin, and K. Spärck Jones, "Effects of out of vocabulary words in spoken document retrieval," in *SIGIR*, Athens, Greece, 2000, pp. 372–374.

[5] A. Yazgan and M. Saraclar, "Hybrid language models for out of vocabulary word detection in large vocabulary conversational speech recognition," in *ICASSP*, Montreal, Canada, 2004, pp. 745–748.

[6] A. Rastrow, A. Sethy, and B. Ramabhadran, "A new method for OOV detection using hybrid word/fragment system," in *ICASSP*, Taipei, Taiwan, 2009, pp. 3953–3956.

[7] T. Schaaf, "Detection of OOV words using generalized word models and a semantic class language model," in *Interspeech*, Aalborg, Denmark, 2001, pp. 2581–2584.

[8] C. Parada, M. Dredze, D. Filimonov, and F. Jelinek, "Contextual information improves OOV detection in speech," in *NAACL-HLT*, Los Angeles, USA, 2010.

[9] J.-L. Gauvain, L. Lamel, and G. Adda, "The LIMSI Broadcast News Transcription System," *Speech Communication*, pp. 89–108, 2002.

[10] J. Feng and A. Sears, "Using confidence scores to improve hands-free speech based navigation in continuous dictation systems," in *ACM Transactions on Computer-Human Interaction*, 2004, pp. 329–356.

[11] J. Lafferty, A. McCallum, and F. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *ICML*, Williamstone, USA, 2001.

[12] J. Fayolle, F. Moreau, C. Raymond, G. Gravier, and P. Gros, "Crf-based combination of contextual features to improve a posteriori word-level confidence measures," in *Interspeech*, Makuhari, Japan, 2010.

[13] G. Damnati and D. Charlet, "Robust speaker turn role labeling of tv broadcast news shows," in *ICASSP*, Prague, Czech Republic, 2011.