



Error detection in confusion network

Alexandre Allauzen

LIMSI/CNRS, Btiment 508, Université Paris-Sud, France

allauzen@limsi.fr

Abstract

In this article, error detection for broadcast news transcription system is addressed in a post-processing stage. We investigate a logistic regression model based on features extracted from confusion networks. This model aims to estimate a confidence score for each confusion set and detect errors. Different kind of knowledge sources are explored such as the confusion set solely, statistical language model, and lexical properties. Impact of the different features are assessed and show the importance of those extracted from the confusion network solely. To enrich our modeling with information about the neighborhood, features of adjacent confusion sets are also added to the vector of features. Finally, a distinct processing of confusion sets is also explored depending on the value of their best posterior probability. To be compared with the standard ASR output, our best system yields to a significant improvement of the classification error rate from 17.2% to 12.3%.

Index Terms: error detection, automatic speech recognition.

1. Introduction

Over the last years, progress has been made in automatic speech recognition (ASR) of broadcast data as a support for an efficient access to audio documents. The NIST evaluations on Spoken Document Retrieval (SDR) showed in 2000 that quality of ASR transcripts turns to be sufficient to enable a variety of applications such as named entities extraction, and content-based document retrieval. The performance of broadcast news (BN) transcription system has been improved since, with reported word error rates in several languages between 10 and 15%.

However, errors in ASR outputs may partially prevent access to data and may simultaneously cause generation of erroneous information. Furthermore, when faced to ill-formed or incorrect sentences further natural language processings become inefficient. To reduce the adverse affects of errors, a solution is to detect errors area in order to, either discard the erroneous part of transcribed speech, or adopt a specific strategy. For example in a SDR system, errors area may be processed with a sub-word based system rather than a word based one.

There are two main approaches for handling ASR errors: filler model and confidence score estimation. On the one hand the use of a filler model involves adding a special lexical entry to the recognizer lexicon. This entry represents out of vocabulary (OOV) words or errors. On the other hand confidence score estimation is based on features estimated by the ASR system. A comparison between these methods showed that confidence score based methods achieve better performance for error detection task [1]. Recent works on confidence score proposed different kind of features extracted during the decoding step, and investigate the use of random forest for confidence estimation [2], SVM regression [3] or neural approach [4] for compensating word posterior estimation bias.

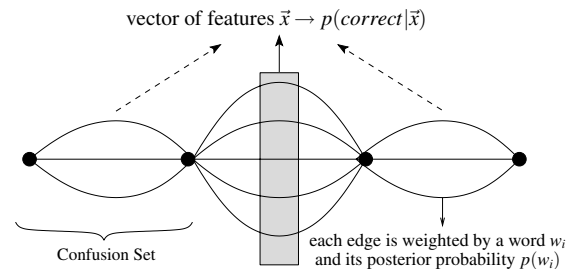


Figure 1: Diagram of a confusion network, and illustration of the feature extraction process for error detection.

In this paper, we explore error detection in a post-processing stage, independently from the ASR system. As shown in figure 1, we propose to make use of features extracted from the confusion networks which is a compact representation of the search space of the ASR system. Inspired by [5], a confidence score is estimated by a logistic regression based on features extracted from confusion sets. Including features from adjacent confusion sets is also investigated as an attempt to involve the neighborhood in the regression process. The features are grouped and assessed according to their type: features solely extracted from the confusion network, those estimated with a language model or from lexical properties of words. As proposed in [3], confusion sets are differently processed whether the posterior probability assigned to the hypothesized word is greater than 0.95 or not. Experiments use the French BN corpus provided with the ESTER evaluation and the real-time transcribing system developed during this evaluation. Results are reported in terms of classification error rate.

In the following, the choice of confusion networks as input data is discussed and features for error modelling are presented. In section 3, the regression model for error detection is defined with the related evaluation framework. Then the corpus and the ASR system are described in section 4, and finally experiments are reported in section 5.

2. Features of confusion

The speech decoding procedure attempts to maximize the posterior probability of the word transcription given the speech signal and statistical models of the speech production. Posterior probabilities are estimated during the conversion of a word lattice into a confusion network [6]. As represented in figure 1, a confusion network is a compact and linear graph, where the complexity of the lattice is reduced to a serie of confusion sets. Each set represents time-parallel word hypotheses with associated posterior probabilities. The words with the highest posterior in each confusion set are hypothesized.

2.1. ASR outputs

Different kind of ASR outputs can be considered as input for confidence estimation. The one-best hypothesis is the most obvious, and the associated word posterior probability is the natural confidence score provided by the ASR system. However, these word posteriors may be over confident. Indeed, recognizers prune the set of explored hypotheses for tractability purpose. Word likelihoods are therefore normalized by a subset rather than the total hypothesis space. Moreover, the possibility of a OOV word yields to an incomplete hypothesis space. Thus, the word posterior probability may be one of the most important feature for confidence score estimation, but not the only one.

In this work, we choose to use confusion networks as input for ASR error detection. This representation contains more information than the one-best hypothesis and without the redundancy of the n -best lists. To be compared with word lattice, the confusion network is more compact and efficient to handle, since the posterior estimate gathered not only words in the best path but also words in competing paths. Thus similar features as described in [7] can be estimated from these both types of graph. However, the conversion from lattice to confusion network merge acoustic and linguistic likelihoods, making features such as acoustic stability no longer accessible.

2.2. Features

As depicted in figure 1, a vector of features \vec{x} is associated to each confusion set (CS). Three feature types are defined according to the involved knowledge sources: the confusion set solely, a statistical language model, and lexical word properties.

The "CS" features are directly derived from the confusion set itself: best posterior probability, duration of the hypothesized word, length of the hypothesized word in number of characters, number of parallel edges, and local entropy of the posterior distribution. The local entropy is defined as:

$$H = - \sum_{i=1}^N p(w_i) \log(p(w_i))$$

where N is the number of competing hypotheses and $p(w_i)$ denotes the posterior of the i -th word of the CS. Information about the context of the confusion set is also added: two boolean features indicating whether the adjacent set (to the left and right) has the null edge¹ as the most probable word, and the length of the confusion network in number of confusion sets.

The language model (LM) features rely on probabilities estimated with a n -gram LM on the confusion network: n -gram likelihood of the hypothesized word; the best LM score when considering all the n -grams derived from the confusion set and its $n - 1$ previous set; the sum of the probabilities of all possible n -grams in the confusion network that can predict the hypothesized word, and the sum of LM scores for all n -grams that can predict the confusion set.

The use of lexical features is also investigated. This features depends on lexical information about the hypothesized word of the CS: the unigram probability, the lexical rank of the word², the number homophones given a lexicon, and the number of possible part-of speech tags for the hypothesized word

¹During the transformation of a word lattice into a confusion network, each word hypothesis is mapped to a position for alignment purpose. Deletion are figured by a null edge.

²The lexical rank of a word is obtained by sorting the words of the vocabulary in a decreasing order according to their frequencies.

(POS-tag ambiguity). This kind of information requires external knowledge sources. The lexical resources of the ASR system described in section 4 provide the unigram distribution over the vocabulary words and the pronunciation lexicon. To estimate the POS-tag ambiguity, we use a POS-tagged version of the newspaper corpus also described in the section 4.

3. Error detection

Confidence score estimation aims to estimate the probability that the best hypothesis of a CS is correct given its associated vector of features \vec{x} . Error detection is a binary and simpler classification task: assign to \vec{x} a class whether the hypothesized word is correct or not. In our modeling, error detection is derived from the estimation of the correctness probability by applying a decision threshold.

3.1. Model for correctness probability

Although there are a variety of types of models that might be used to relate the vector \vec{x} to the target probability, we explore the use of generalized linear models (glm) for confidence estimation from confusion network. This kind of models were introduced by [5] for confidence estimation from the one-best hypothesis. The glm assumes that

$$g(p) = \vec{a} \cdot \vec{x} + b,$$

where p is the target confidence or correctness probability, g is a monotone function, called the link function. This function aims to map the unit interval to the real line. The vector \vec{a} and the "intercept" term b are the unknown parameters of this model. In this work, the link function is the most commonly used *logit* function $g(p) = \log(p/(1-p))$, which yields to the well known logistic regression. The parameters \vec{a} and b are estimated under the maximum likelihood criterion.

Given the correctness probability p , error detection is performed by applying a threshold on p . This threshold may be tuned to weight the cost of false alarm *versus* missed detection errors. In the following experiments, the default value for this threshold is set to 0.5.

3.2. Evaluation

Results are evaluated using the classification error rate (CER). To set up the baseline result, we may assume that the most probable words given by the confusion network are always correct. This leads to the baseline CER which is equal to the word error rate (WER). These two measures are defined as follow:

$$\begin{aligned} CER &= \frac{\#(CS \text{ correctly classified})}{\#(CS)} \\ WER &= \frac{\#(substitution+insertion)}{\#(hypothesized \ words)} \end{aligned}$$

The WER does not include as usual errors of deletion and is normalized by the number of words of the hypothesis, since our goal is to classify CSs associated to each hypothesized word.

4. Corpus and ASR system

The following experiments use as data set the official test of the French ESTER campaign. This corpus consists of 10 hours of French speaking radio broadcast news shows, taken from four different stations occurring in the official training data of

<i>Data set</i>	<i>Total</i>	<i>Errors</i>	<i>Error rate</i>
All data	107602	18472	17.2
$post \leq 0.95$	57823	16150	27.9
$post > 0.95$	49779	2322	4.7

Table 1: Summary of the generated data in terms of number of CS and error rates. Two sub-categories are distinguished, whether the best posterior probability of the CS (*post*) is lower or greater than 0.95.

the evaluation, one source (France Culture) without any transcribed training data and one unknown source. This test set was recorded from October to December 2004 and contains 107k words after normalization.³

The LIMSI French real-time transcribing system [8] is used to generate the confusion networks. The vocabulary contains 65k words and the language models are obtained by the linear interpolation of four standard n -gram back-off LM respectively trained on about 3.7M words of precise transcriptions of BN acoustic data, 89M words of rapid transcriptions of BN data, 530M words of newspaper and newswire data and the newspaper texts (370M words) distributed for the ESTER evaluation. These LMs are also used to estimate some features. The acoustic models were trained on about 190 hours of BN training data. And the speech features consist of 39 cepstral parameters derived from a Mel frequency spectrum estimated on the 0-8kHz band (or 0-3.8kHz for telephone data) every 10ms. The decoding consists in two passes with bigram LM, very tight pruning thresholds (especially for the first pass) and fast Gaussian computation based on Gaussian short lists. Each pass generates a word lattice which is expanded with a 4-gram LM. This expanded lattice is finally converted to a confusion network. More details on the system and the training data can be found in [8].

5. Experiments

The reported results are obtained using the *R-project* tools⁴ to estimate the regression model, and the *SRI-LM* toolkit⁵ to estimate and handle LMs. The confusion networks are generated for the whole ESTER test corpus and aligned with the reference transcript using the standard dynamic programming algorithm. If *ex-æquo* alignments occur, a constraint on Levenshtein distance between words is applied to select one alignment among the other. Before alignment, normalization rules are usually applied on the word hypothesis like case-insensitive and compound words conversions. In this work, these rules are modified to preserve the word segmentation of the ASR system. The baseline WER is therefore higher than the official one.

For error detection experiments, a tag is associated to each CS whether the hypothesized word is correct or is an error. Although they are used to estimate some features, the confusion sets with the null edge as most probable hypothesis are not considered for the error detection task. A corpus of 107k tagged CS is thus obtained, and the overall characteristics of the data and their partitions are summarized in table 1. It can be observed that applying a threshold of 0.95 to the best posterior probability (*post*) divides the data in two partitions of comparable size but with different error rates. To provide two test corpus, 10k

³More details on data sets, systems and the results can be found on the web site <http://www.afcp-parole.org/ester/index.html>

⁴<http://www.r-project.org/>

⁵<http://www.speech.sri.com/projects/srilm/>

<i>Feature type</i>	<i>CER</i>
Baseline system	27.9
LM	27.7
Lexical	27.5
Contextual	26.8
Confusion set	20.9
Confusion set + contextual	20.4
All	20.3
Adjacent (confusion set +contextual)	20.3
Adjacent (all)	19.9

Table 2: Impact of the different feature types in terms of classification error rate (CER). The feature types are first evaluated independently in the upper part, then different combinations are assessed. Results are given on data with $post \leq 0.95$. "Adjacent" means that the considered features from adjacent CS are added to the vector of features for the CS to be classified.

<i>Feature</i>	<i>t-value</i>
best posterior probability	0.487
local entropy	0.296
duration of the hypothesized word	0.115
number of parallel edges	0.045
log length of the confusion network	0.015
null edge on the left	0.015
null edge on the right	0.013
length of the hypothesized word	0.007

Table 3: The *t-values* associated to each feature for the combination of the CS and contextual features.

CSs are randomly sampled from these two partitions. The rest of the data provided two training sets.

5.1. Impact of feature sets

Results obtained with various combination of feature types are reported in table 2 on data with $post \leq 0.95$. Each feature type is first evaluated independently. These types are defined in subsection 2.2. As shown by the upper part of table 2, the use of features based on language model or lexical resources does not result in a great improvement. These results are obtained with the same bigram LM used by the ASR system, and experiments with a trigram LM do not change significantly the results. We may assume that informations carried by these models are already used by the speech decoder.

On the contrary, the use of features solely extracted from the confusion networks yields to a significant improvement. The CER is reduced by a quarter when using only the type of features referred as "confusion set" in table 2: the best posterior probability, the local entropy, the number of parallel edges, the duration and the length of the hypothesized word. This relative gain raises to 27% by adding the "contextual" features (null edge as hypothesized word in the left or right CS, and the length of the confusion network). The impact of each feature may be sorted by ranking their associated *t-values*. This measure is the ratio of the estimated parameter value to its standard error. The *t-values* are reported in table 3 for the combination of CS and contextual feature types. It can be observed that the greatest predictor is the best posterior probability as expected. The following predictors are in decreasing order, the local entropy, duration, and the number of parallel edges. The impact of other features are less significant.

Joining together all the features does not outperform the combination of CS and contextual features. To involve the neighborhood in the regression process, the features of adjacent CSs are added in \vec{x} , multiplying its dimension by three. This yields to the best performance with a CER of 19.9%, which represents a relative reduction of 28%.

5.2. Impact of data partitioning

A cross evaluation is performed to assess the impact of partitioning the data depending on *post*: two logistic regression models are estimated on each training corpus and tested on both test corpus. Results are given in table 4. When looking at the first column, it can be inferred that training the classifier on data with $post > 0.95$ to classify data with $post \leq 0.95$ performs poorly to be compared with the classifier trained on data with $post \leq 0.95$ (a CER of 29.1% vs 19.9%). Nevertheless, merging all training data to classify data with $post \leq 0.95$ does not impoverish the CER. It can also be observed that whatever the training set, the CER still quite identical to the baseline one, when testing on data with $post > 0.95$. Thus an appropriate decision rule to classify this kind of CS may be simply to decide that the ASR system is still right. These results confirm the assumption of [3] and justify the partitioning of the data.

5.3. The best system

Our best system for confidence estimation and error detection is obtained as follow: the training data corresponding in a $post \leq 0.95$ are used to estimate a logistic regression model using all the features described in this article, including the features from adjacent sets. This model is used to estimate the correctness probability of CSs with a $post \leq 0.95$. For the other CSs, the ASR system is assumed to always hypothesize the right word. With this approach, our system achieve a CER of 12.3% on the whole test set of 20k confusion sets. This represents a relative reduction of 28.5%.

6. Conclusions

In this article, we addressed the problem of error detection for a BN transcription system in a post-processing stage. To estimate a correctness probability we proposed to use logistic regression based on features extracted from confusion networks. Error detection is then achieved by applying a decision threshold on this probability. Different kinds of knowledge sources were investigated such as the confusion set solely, language model, and lexical word properties. Experiments were carried out on the French test set of ESTER and the evaluation was performed using the classification error rates. Results showed that the features extracted from the confusion network only and its properties have the greatest impact on performances. In particular, by examining the *t-values* associated to each feature, we observed that, besides the posterior probability of the hypothesized word, the most important features are: the local entropy, the duration of the hypothesized word, the number of parallel edges, the length of the confusion network, the presence of null as hypothesized word in the left and the right confusion set. With these features, the CER is reduced by 27% relative.

To enrich our modeling with information about the neighborhood, features of adjacent confusion sets was included in the vector of features and leads to a slight improvement. Furthermore, we assessed the distinct processing of confusion sets depending on the value of the best posterior probability. A logistic regression model is trained to estimate the confidence score of

Training set	Test set	
	post \leq 0.95	post $>$ 0.95
post \leq 0.95	19.9	4.9
post $>$ 0.95	29.1	4.5

Table 4: Cross evaluation with the different data partitions: for each part of the training data a regression model is trained and evaluated on each test corpus.

confusion sets with best posterior lower than 0.95, while for the other confusion sets the best posterior given by the ASR system is considered as the confidence score. With this approach, the classification error rate is reduced from 17.2% to 12.3%. In future works, we plan to address error classification issues and more specifically OOV detection.

7. Acknowledgments

This work was partially funded by the European Union under the integrated project TC-STAR (IST-2002-FP6-506738), and the French government under the *Cap Digital* project *Infom@gic*. The author wish to thank Mohamed Ben Diaby for his contribution to this work.

8. References

- [1] T. Hazen and I. Bazzi, "A comparison and combination of methods for OOV word detection and word confidence scoring," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, May 2001.
- [2] J. Xue and Y. Zhao, "Random forests-based confidence annotation using novel features from confusion network," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006.
- [3] D. Hillard and M. Ostendorf, "Compensating for word posterior estimation bias in confusion networks," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, Toulouse, France, May 2006.
- [4] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, and A. Stolcke, "Neural-network based measures of confidence for word recognition," in *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997, pp. 887–890.
- [5] L. Gillick, Y. Ito, and J. Young, "A probabilistic approach to confidence estimation and evaluation," in *Proceedings of the 1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, Munich, Germany, 1997, pp. 879–882.
- [6] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: Word error minimization and other applications of confusion networks," *Computer, Speech and Language*, vol. 14, no. 4, pp. 373–400, 2000.
- [7] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," in *Proc. Eurospeech '97*, Rhodes, Greece, 1997, pp. 827–830. [Online]. Available: cite-seer.ist.psu.edu/kemp97estimating.html
- [8] J.-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, and H. Schwenk, "Where Are We in Transcribing French Broadcast News?" in *InterSpeech*, Lisbon, September 2005.