Improving ASR error detection with non-decoder based features

Thomas Pellegrini[1], *Isabel Trancoso*[12]

[1]INESC-ID Lisboa, Portugal
[2]IST, Lisboa, Portugal
thomas.pellegrini@l2f.inesc-id.pt

Abstract

This study reports error detection experiments in large vocabulary automatic speech recognition (ASR) systems, by using statistical classifiers. We explored new features gathered from other knowledge sources than the decoder itself: a binary feature that compares outputs from two different ASR systems (word by word), a feature based on the number of hits of the hypothesized bigrams, obtained by queries entered into a very popular Web search engine, and finally a feature related to automatically infered topics at sentence and word levels. Experiments were conducted on a European Portuguese broadcast news corpus. The combination of baseline decoder-based features and two of these additional features led to significant improvements, from 13.87% to 12.16% classification error rate (CER) with a maximum entropy model, and from 14.01% to 12.39% CER with linear-chain conditional random fields, comparing to a baseline using only decoder-based features.

Index Terms: error detection, automatic speech recognition.

1. Introduction

Error detection is an important topic in Automatic Speech Recognition (ASR). Three types of errors can occur in the hypothesized word stream output: substitutions, insertions and deletions. Having a confidence measure indicating a potential substitution or insertion error for each hypothesized word is useful in many applications, for instance to discard sentences with errors in real-time broadcast news (BN) subtitling systems. Also, the ability to label words recognized with low confidence in an automatically recognized transcript is very relevant for our computer aided language learning system, where BN videos with automatically produced captions may make the use of the system more motivating for students [1].

Error detection can be performed by making a binary decision with the help of a statistical classifier, but most often a probability of how reliable is a hypothesized word is first estimated, and then a decision is made by using a threshold on this probability called a confidence measure. There are at least two main approaches to estimate confidence measures: by directly estimating the posterior probability of the hypothesized word, or by using predictor features collected during decoding [2]. The first approach requires the estimation of a filler model in order to compute an "all event" probability needed to normalize the posterior score given by the ASR decoder. The second approach is much easier to perform, since information comes from the decoder. Nevertheless, there is no ideal predictor feature. The overlap between correct and wrong hypothesized words is large, even for the best predictor feature. In this study, we explored new features, coming from other knowledge sources to bring additional information, complementary to the information provided by decoder-based features. In the literature, prosodic cues for example, like pitch excursion, loudness, prior pause and overall duration for user turns, have shown to improve significantly misrecognition prediction [3]. Other cues were explored in this study: word match feature between two system outputs, number of bigram hits by querying a popular Web search engine, and finally automatic topic detection.

Many statistical tools have been proposed in the literature to estimate confidence measures, for instance: generalized linear models [4], artificial neural networks [5], maximum entropy models [6], and more recently linear-chain conditional random fields [7]. The last two techniques have been chosen in this study, for their discriminative capabilities.

2. Models for error detection

Many distinct types of statistical classifiers can be used. Currently, our in-house ASR system estimates confidence measures with a maximum entropy model (Maxent). In this study, we compared this technique to linear-chain conditional random fields. Both models were used as the following: when the probability or confidence measure given by the model is lower than 0.5, then the hypothesized word is labeled as an error.

2.1. Maximum entropy models

Maximum Entropy models are very popular discriminant models, and are used in many applications, in particular in natural language processing tasks, such as part-of-speech tagging. The Maxent principle states that the correct probability distribution for a given class is the one that maximizes entropy, given constraints on the distribution. One advantage of Maxent models is that the training algorithm will determine how to combine the different features by estimating the best feature weights, so that the main user effort will consist of identifying which features are best to be used. To train and infer the Maxent model, we used the Megam toolbox, available at `http://www.cs.utah.edu/~hal/megam`.

2.2. Linear-chain conditional random fields

Introduced by Laferty et al (2001), conditional random fields (CRF) are also discriminant models. The conditional distribution of the labels that we wish to predict, given feature observations, is associated to a graphical structure, that allows to model complex dependencies between output and input variables. When then output variables are arranged in a sequence, the graphical structure is a linear chain, and in this particular case, CRFs are called linear-chain CRFs. Linear-chain CRFs are often presented as a sequential version of Maxent models, and a discriminant version of Hidden Markov Models [8]. We used the Java-based Mallet package [9].

3. Baseline features

The output of the ASR system is a stream of words. For each hypothesized word, various decoder-based features are available. In this study, only words from the best hypothesis are considered. Our baseline classifier is based on the following set of 15 features, commonly used in ASR error detection:

- Global, acoustic and posterior scores,
- Average phone acoustic and posterior scores,
- Length of words in number of decoding frames (20 ms duration) and in number of phones,
- Log of the total and average active states, arcs and tokens,
- Minimum and average phone log-likelihood ratios.

Features related to the active states, arcs and tokens for each hypothesized word should intuitively have large values to reflect a large degree of uncertainty of the recognizer [10].

4. New features for error detection

4.1. Binary word match feature

In [11, 12], outputs from two different systems were used to improve confidence measure (CM) estimation or OOV/error detection: a strongly constrained recognizer with a word-based language model, and a weakly constrained phone-based recognizer. In [11], words and phones from the two outputs are aligned and compared. In [12], CMs from the two systems are combined to improve their estimation. If the two systems produce inconsistent output for a given speech segment, hence perhaps the strong system has made an error. In both cases, a phone-to-word transducer is needed. We propose to explore a similar idea, with a simpler approach that does not need to build a new transducer nor a phonotactic language model. In our approach, the two ASR systems differ only on the Acoustic Model (AM) sets: a context-dependent acoustic model set (our strong system), and a monophone set. To build context dependent AMs, monophone models are first needed, hence no new system has to be built. Since the two systems are close in performance, most of the errors occur in the same speech segments, but the errors are different. Hence, words in the same position that differ from both transcripts are probably errors. The comparison feature, named hereafter "w", is a binary word match feature that compares the two outputs at word level. For performance comparisons between monophone and context dependent units based ASR systems, the reader may refer to [13].

An example of how w is computed, is illustrated in table 1. ASR1 and ASR2 are the outputs of respectively our best system, and the monophone-based system. The corresponding excerpt of the manually transcribed sentence was: "*ataques ao afeganistão washington diz que os.*" The alignment of ASR1 and ASR2, and the feature values are indicated respectively in the third and fourth rows. The word "ojdanic" was found to be substituted by "os", and consequently was assigned a feature value of 0. Indeed this word was misrecognized, since it would correspond to "washington diz que" in the reference.

4.2. Bigram hit feature

The second new feature was the number of hits found by querying a very popular Web search engine, at bigram level. For a given hypothesized sentence, two hit values per word were retrieved and used as features: one with the preceding word and one with the next word. Only one hit feature value was computed for the first and last word of a sentence, since there were respectively no preceding word or next word for these two sentence boundary words. Queries were performed by surrounding the bigrams by quotes to force the search engine to retrieve the bigrams as is.

ASR1:	ataques	ao	afeganistão	ojdanic			os
ASR2:	ataques	ao	afeganistão	os	danos	que	os
	OK	OK	OK	SUBS	INS	INS	OK
w value	1	1	1	0			1

Table 1: Word binary match feature 'w' example. The third line is a word-by-word comparison between the two ASR outputs, with these abbreviations: SUBS: substitution, INS: insertion.

Raw scaled hit values did not show improvement, but quantized values did. The hit values were quantized between 0 and 1, using simple heuristic rules, given in table 2. Raw hit values and quantized feature values are given in table 3, for our previous example excerpt. The sentence beginning bigram "ataques ao" query resulted in 584k hit values, which was quantized as a maximum feature value of one. The two bigrams involving the misrecognized word "ojdanic" led to very low hit values. In general, a bigram showing a zero hit value has very likely one or both of its words misrecognized.

This feature could have been computed from the language model from our ASR system, but the use of a Web search engine seemed advantageous, because of its evergrowing and up-to-date indexed content.

$h = 0 \rightarrow h = 0.0$
$h > 0$ and $h < 101 \rightarrow h = 0.2$
$h > 100$ and $h < 1001 \rightarrow h = 0.4$
$h > 1000$ and $h < 10001 \rightarrow h = 0.6$
$h > 10000$ and $h < 100001 \rightarrow h = 0.8$
$h > 100000 \rightarrow h = 1.0$

Table 2: Heuristic rules used to quantize raw hit values 'h'.

4.3. Topic feature

Very often, misrecognitions appear to be out of the global topic of the hypothesized sentence. Hence, a feature that would estimate how much a word is related to a given topic, in this case the topic of the document being transcribed, or the topic of the sentence to which the word belongs to, is expected to help detecting errors.

Since our corpus is not labeled in terms of topics, and also for generalization purpose, unsupervised topic models were

	Raw bigram hits		Feature values	
ataques	_	584k	_	1.0
ao	584k	255k	1.0	1.0
afeganistão	255k	0	1.0	0.0
ojdanic	0	22	0.0	0.2
os	22	_	0.2	_

Table 3: Bigram hit feature example. *Second and third columns*: raw hit values of the two bigrams per word. *Fourth and fith columns*: corresponding quantized values.

Train		*Test*	
#Words	108,029	*#Words*	16,518
Errors	*Correct words*	*Errors*	*Correct words*
14,542	93,487	2,579	13,939

Table 4: *Number of errors (misrecognitions) and correct words in train and test sets. Errors were considered as the positive class.*

needed. In [14], Latent Semantic Analysis was used to define a semantic similarity between words, to derive confidence measures. Here, we used again the Mallet package to train and infer topic models using the Gibbs sampling approach, for which a topic consists of a cluster of words that frequently occur together [9]. The number of topics has to be chosen, depending on the application. Several numbers were tested, from 10 topics, to provide a broad overview of the contents of the corpus, to 300 topics, to obtain fine-grained results. Results with different number of topics were similar. Experiments with 100 topics are reported hereafter.

Topic models were trained on the manual transcriptions of our training corpus and a corpus of 31 million words, comprised of texts published in a Portuguese national newspaper during 2001, the time period of our test corpus. Topic models provide a vector of topic weights for all the words of the training corpus. Once topic models were trained, topics were infered for the ASR outputs at sentence-level. A vector of weights was infered for each ASR hypothesized sentence, at sentence-level, and compared to vectors at word-level. When a hypothesized word was not seen in the training corpus, no feature value was attributed to the word. A cosine similarity measure with values in $[0, 1]$, common in information retrieval, was used: $s = \mathbf{x}.\mathbf{y}/||\mathbf{x}||.||\mathbf{y}||$. A large similarity value between topic vectors of a hypothesized word and the sentence it belongs to, was expected to increase the confidence in this word.

5. Corpus

The corpus used in this study is a subset of the ALERT European Portuguese BN corpus [15]. It is comprised of 14 manually transcribed TV newscasts, recorded in 2001, totalizing 14 hours of speech. The two most recent newscasts were used for test.

All the material has been automatically transcribed by using our in-house speech recognition system. Word error rate for the test subset was 15.61%. Automatic transcriptions were aligned with the corresponding manual transcriptions to provide material to train and test the misrecognition classifiers. Table 4 gives the number of automatically transcribed words for the train and test subsets. The numbers of correctly recognized words and errors are also indicated.

6. Experiments

Automatic transcriptions were performed with our in-house ASR system, named AUDIMUS, a hybrid Artificial Neural Networks / Hidden Markov Models system [16]. Our strong system used a set of 385 context dependent diphone-like acoustic models (AMs), trained on about 1,000 hours of broadcast news speech. To compute the w feature, a monophone-based weaker system was used, with a set of 40 phonemes. More details about the context dependency modelling and performance comparison can be found in [13]. Only the AMs differ in both systems.

	CER(%)	minDCF (x10^{-2})
M	13.87	23.08
Mw	12.31	19.74
Mh	13.49	22.60
Mt	13.83	23.07
Mwh	12.16	19.74
C	14.01	22.34
Cw	12.39	20.16
Ch	13.79	22.20
Ct	14.11	22.49
Cwh	12.46	20.83

Table 5: *Classification error rates (CER) and minima of the Detection Cost Functions (minDCF) for the various feature sets. Upper part: Maxent 'M', lower part: CRF 'C'.*

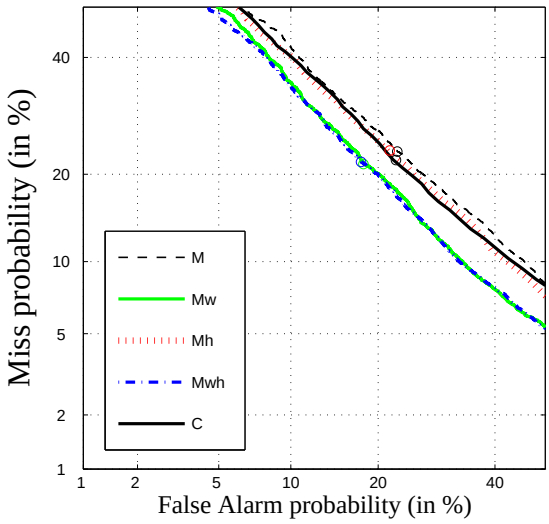

Figure 1: DET curves, for the most relevant feature sets used with M, and baseline feature set with C. Other DET curves for C are similar.

Errors can be detected only on the hypothesized words, thus only substitutions and insertions are addressed, not deletions. Error detection was evaluated with a global Classification Error Rate (CER), defined as the ratio of the number of misclassifications over the number of hypothesized words. To complete the information given by CERs, we also present Detection Error Trade-off (DET) curves and associated minimum Detection Cost Function rates (DCF), which plot the False Alarm (FA, impostor attempts accepted) probability as a function of the Miss or False Rejections (FR, genuine attempts rejected) probability. DET curves are a standard means of representing performance on detection tasks, to help comparing systems according both to error rates. DCFs are a weighted sum of the FA and FR rates. In our experiments, both weights were chosen equal. Errors were labeled as the positive class, since we were interested in detecting errors.

6.1. Results

Figure 1 shows DET curves achieved on the test data. For clarity, only the most relevant DET curves were plotted in the fig-

ure. Table 5 presents the corresponding classification results. The table shows the CERs and minimum DCFs, for the different feature sets and the two classifiers. In both the figure and table, M denotes the Maxent model and C the linear-chain CRF.

Baseline performances correspond to the M and C rows in the table, with the 15 decoder-based feature set. The CRF baseline slightly outperformed the Maxent baseline, with a minDCF reduction from 23.08% to 22.34%. The C DET curve clearly shows better performance than the M curve for the complete range of operating points. Nevertheless, the C CER is larger, due to the fact that CMs computed with CRFs are larger in average than those computed with Maxent.

In the table, lines with a gray background color show the best performances achieved. Adding the w feature to the baseline feature set led to the largest and most significative improvements, given in the Mw and Cw rows: for minDCFs, 15% relative reduction from 23.08% to 19.74% for M, and 10% relative from 22.34% to 20.16% for C. For M, the best CER was achieved by using the w and h additional features (Mwh row), 12.3% relative reduction from 13.87% for the baseline to 12.16%.

The hit feature h led to slight improvements for both M and C, from 13.87% to 13.49% for M CER, and from 14.01% to 13.79% for C CER. Using both h and w features did not improve the C performance, but gave the best results with the M classifier. Mwh was actually the best classifier, over all feature sets and in comparison to all the C classifiers.

The topic feature t did not show any improvement in comparison to the baseline. It may be explained by the difficulty to use unsupervised topic models at word level, more often used at document level. More work is needed to fully explore this feature.

7. Summary and future work

In order to improve error detection in ASR output, we explored features coming from different existing knowledge sources to complement baseline decoder-based features. Two out of three new features, a binary word match feature and a bigram hit feature, led to significant improvements, from 13.87% to 12.16% CER with a maximum entropy model, and from 14.01% to 12.39% CER with linear-chain conditional random fields, comparing to a baseline using only decoder-based features. The third feature related to automatically infered topics at sentence and word levels did not show improvement.

Since the purpose was to test new features, experiments were conducted only on one-best hypothesis. We plan to validate our approach by extending the baseline feature set with word lattice or confusion network features, like local entropy to take into account competing word hypothesis. More experiments are needed to test topic-based features, and in general semantic-based features.

8. Acknowledgments

This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds and by FCT project CMU-PT/HuMach/0039/2008. The authors would like to thank Alexandre Allauzen, from LIMSI-CNRS, for his help in the use of the CRF Mallet package.

9. References

[1] L. Marujo, J. Lopes, N. Mamede, I. Trancoso, J. Pino, M. Eskenazi, J. Baptista, and C. Viana, "Porting REAP to European Portuguese," in *proceedings of SLATE 2009 - Speech and Language Technology in Education*, Brighton, 2009.

[2] H. Jiang, "Abstract confidence measures for speech recognition: A survey," *Speech Communication*, vol. 45, pp. 455–470, 2004.

[3] J. Hirschberg, D. Litman, and M. Swerts, "Prosodic and other cues to speech recognition failures," *Speech Communication*, vol. 43:1-2, pp. 155–175, 2004.

[4] A. Allauzen, "Error detection in confusion network," in *proceedings of INTERSPEECH*, Antwerp, 2007, pp. 1749–1752.

[5] M. Weintraub, F. Beaufays, Z. Rivlin, Y. Konig, and A. Stolcke, "Neural - Network Based Measures of Confidence for Word Recognition," in *proceedings of ICASSP*, Munich, 1997, pp. 887–890.

[6] C. White, J. Droppo, A. Acero, and J. Odell, "Maximum entropy confidence estimation for speech recognition," in *proceedings of ICASSP*, Hawaii, 2007, pp. 809–812.

[7] J. Xue and Y. Zhao, "Random forests-based confidence annotation using novel features from confusion network," in *proceedings of ICASSP*, Toulouse, 2006, pp. 1149–1152.

[8] C. Sutton and A. McCallum, *An Introduction to Conditional Random Fields for Relational Learning, In Lise Getoor and Ben Taskar, editors. Introduction to Statistical Relational Learning.* MIT Press, 2007.

[9] C. Sutton, *GRMM: GRaphical Models in Mallet*, 2006. [Online]. Available: http://mallet.cs.umass.edu/grmm/

[10] L. Gillick, Y. Ito, and J. Young, "A probabilistic approach to confidence estimation and evaluation," in *proceedings of ICASSP*, Munich, 1997, pp. 879–882.

[11] C. White, G. Zweig, L. Burget, P. Schwarz, and H. Hermansky, "Confidence Estimation, OOV Detection, and Language ID using Phone-to-Word Transduction and Phone-Level Alignments," in *proceedings of ICASSP*, Las Vegas, 2008, pp. 4085–4088.

[12] L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C. White, S. Khudanpur, H. Hermansky, and J. Cernocky, "Combination of Strongly and Weakly Constrained Recognizers for Reliable Detection of OOVs," in *proceedings of ICASSP*, Las Vegas, 2008, pp. 4081–4084.

[13] A. Abad and J. Neto, "Incorporating Acoustical Modelling of Phone Transitions in a Hybrid ANN/HMM Speech Recognizer," in *proceedings of INTERSPEECH*, Brisbane, 2008, pp. 2394–2397.

[14] S. Cox and S. Dasmahapatra, "High-level Approaches to Confidence Estimation in Speech Recognition," *IEEE Trans. Speech and Audio Processing*, vol. 10:7, pp. 460–471, 2002.

[15] R. Amaral, H. Meinedo, D. Caseiro, I. Trancoso, and J. Neto, "A Prototype System for Selective Dissemination of Broadcast News in European Portuguese," *EURASIP Journal on Advances in Signal Processing*, vol. 37507, 2007.

[16] H. Meinedo, D. Caseiro, J. Neto, and I. Trancoso, "AUDIMUS.media: a broadcast news speech recognition system for the european portuguese language," in *proceedings of PROPOR*, Faro, 2003, pp. 9–17.