Reference-free and Confidence-independent Binary Quality Estimation for Automatic Speech Recognition

Hamed Zamani[⋆], José G. C. de Souza[†‡], Matteo Negri[‡], Marco Turchi[‡], Daniele Falavigna[‡]
[⋆]School of ECE, College of Engineering, University of Tehran, Iran
[†]University of Trento, Italy
[‡]HLT research unit, Fondazione Bruno Kessler, Trento, Italy
h.zamani@ut.ac.ir {desouza,negri,turchi,falavi}@fbk.eu

Abstract

We address the problem of assigning binary quality labels to automatically transcribed utterances when neither reference transcripts nor information about the decoding process are accessible. Our quality estimation models are evaluated in a large vocabulary continuous speech recognition setting (the transcription of English TED talks). In this setting, we apply different learning algorithms and strategies and measure performance in two testing conditions characterized by different distributions of "good" and "bad" instances. The positive results of our experiments pave the way towards the use of binary estimators of ASR output quality in a number of application scenarios.

1 Introduction

Accurate and cost-effective methods to estimate ASR output quality are becoming a critical need for a variety of applications, such as the large vocabulary continuous speech recognition systems used to transcribe audio recordings from different sources (*e.g.* YouTube videos, TV programs, corporate meetings), or the dialogue systems for human-machine interaction. For obvious efficiency reasons, in some of these application scenarios, ASR output quality cannot be determined by means of standard reference-based methods. Indeed, besides the fact that reference transcripts are not always available, quality indicators should often be computed at run-time to ensure quick response. This motivates research towards alternative **"reference-free"** solutions. To cope with this problem, word-level confidence estimates have been used in the past either to measure how an ASR system is certain about the quality of its hypotheses (Wessel et al., 1998; Evermann and Woodland, 2000; Mangu, 2000; Xu et al., 2010, inter alia) or to automatically detect ASR errors (Seigel, 2013; Seigel and Woodland, 2014; Tam et al., 2014). The reliance on confidence information and the emphasis on the word/sub-word level mark the major differences between such prior works and our research, which aims to give an objective assessment of ASR output quality: *i)* at the whole utterance level and *ii)* without the constraint of having access to the system's decoding process. This information, in fact, is not always accessible, as in the case of the increasingly large amount of captioned audio/video recordings that can be found on the Web. This advocates for the development of **"confidence-independent"** quality estimation methods.

These problems have been addressed by Negri et al. (2014), who proposed the task of predicting the word error rate (WER) of an automatically transcribed utterance.[1] Results indicate that even with a relatively small set of *black-box* features (*i.e.* agnostic about systems' internal decoding strategies), the predictions closely approximate the true WER scores calculated over reference transcripts. Experiments, however, are limited to a regression problem and further developments either disregard its natural extension to binary classification (Jalalvand et al., 2015), or address it without the same exhaustiveness of this work (C. de Souza et al., 2015).

The automatic assignment of explicit good/bad labels has several practical applications. For instance, instead of leaving to the user the burden of interpreting scores in the [0, 1] interval, easily-interpretable binary quality predictions would help in tasks like: *i)* deciding if an utterance in a dialogue application has been correctly recognized, *ii)* deciding if an automatic transcription is good enough for the corresponding audio recording or needs manual revision (*e.g.* in subtitling applications), *iii)* selecting training data for acoustic modelling based on active learning, and *iv)* retrieving audio data with a desired quality for subsequent processing in media monitoring applications.

To support these applications, we extend ASR

[1]This "quality estimation" task presents several similarities with its counterpart in the machine translation field (Specia et al., 2009; Mehdad et al., 2012; Turchi et al., 2014; C. de Souza et al., 2014, *inter alia*).

quality estimation to the binary classification setting and compare different strategies. All of them significantly outperform the trivial approach based on thresholding predicted regression scores (our first contribution). The best solution, a *stacking method* that effectively exploits the complementarity of different models, achieves impressive accuracy results (our second contribution).

2 Methodology

Task Definition. Given a set of *(signal, transcription, WER)* tuples as training instances, our task is to label unseen *(signal, transcription)* test pairs as "good" or "bad" depending on the quality of the transcription. The boundary between "good" and "bad" is defined according to a threshold τ set on the WER of the instances: those with a $WER \leq \tau$ will be considered as positive examples while the others will be considered as negative ones. Different thresholds can be set in order to experiment with testing conditions that reflect a variety of application-oriented needs. At the two extremes, values of τ close to zero emphasize systems' ability to precisely identify high-quality transcriptions (those with $WER \leq \tau$), while values of τ close to one shift the focus to the ability of isolating the very bad ones (those with $WER > \tau$). In both cases, the resulting datasets will likely be rather imbalanced, which is a challenging condition from the learning perspective.

Approaches. We experiment with two different strategies. The first one, *classification via regression*, represents the easiest way to adapt the method proposed in (Negri et al., 2014). It fits a regression model on the original training instances, applies it to the test data, and finally maps the predicted regression scores into good/bad labels according to τ. The second one is *standard classification*, which partitions the training data into good/bad instances according to τ, trains a binary classifier on such data, and finally applies the learned model on the test set. The two strategies have pros and cons that are worth to consider. On one side, classification via regression directly learns from the WER labels of the training points. In this way, it can effectively model the instances whose WER is far from the threshold τ but, at the same time, it is less effective in classifying the instances with WER values close to τ. Moreover, in case of skewed label distributions, its predictions might be biased towards the average of the training labels. Nevertheless, since such mapping is performed *a posteriori* on the predicted labels, the behaviour of the model can be easily tuned with respect to different user needs by varying the value of τ. On the other side, standard classification learns from binary labels obtained by mapping *a priori* the WER labels into the two classes. This means that the behaviour of the model cannot be tuned with respect to different user needs once the training phase is concluded (to do this, the classifier should be re-trained from scratch). Also, standard classification is subject to biases induced by skewed label distributions, which typically results in predicting the majority class. To cope with this issue, we apply instance weighting (Veropoulos et al., 1999) by assigning to each training instance a weight *w* computed by dividing the total number of training instances by the number of instances belonging to the class of the given utterance.

Since classification via regression and standard classification are potentially complementary strategies, we also investigate the possibility of their joint contribution. To this aim, we experiment with a *stacking method*, or stacked generalization (Wolpert, 1992), which consists in training a meta-classifier on the predictions returned by an ensemble of base classifiers. To do this, training data is divided in two portions. One is used to train the base estimators; the other is used to train the meta-classifier. In the evaluation phase, the base estimators are run on the test set, their predictions are used as the features for the meta-classifier, and its output is returned as the final prediction.

Features. Similar to Negri et al. (2014), we experiment with 68 features that can be categorized into Signal, Hybrid, Textual, and ASR. The first group is extracted by looking at each voice segment as a whole. Hybrid features give a more fine-grained information obtained from knowledge of word time boundaries. Textual features aim to capture the plausibility/fluency of an automatic transcription. Finally, ASR features give information based on the confidence the ASR system has on its output. Henceforth, we will refer to the first three groups as "*black-box*" features since they are agnostic about the system's internal decoding process. The fourth group, instead, will be referred to as the "*glass-box*" group since they consider information about the inner workings of the ASR system that produced the transcriptions. The glass-box features will be exploited in §3 to train the full-fledged quality estimators used as terms of comparison in the evaluation of our confidence-independent models.

To gather insights about the usefulness of our

features, in all our experiments we performed feature selection using Randomized Lasso, or stability selection (Meinshausen and Bhlmann, 2010). Interestingly, the selected black-box features are uniformly distributed in all the groups; this suggests to keep all of them (and possibly add others, which is left for future work) while coping with binary quality estimation for ASR.

Learning Algorithms. Besides comparing the results achieved by different learning strategies, we also investigate the contribution of various widely used algorithms. For *classification via regression* we use Extremely Randomized Trees (XTR (Geurts et al., 2006)) and Support Vector Machines (SVR (Cortes and Vapnik, 1995)) regressors, while for *standard classification* we use Extremely Randomized Trees (XTC), Support Vector Machine (SVC (Mammone et al., 2009)), and Maximum Entropy (MaxEnt (Csiszár, 1996)) classifiers. MaxEnt is also used as the meta-classifier by our *stacking method*. In all experiments, hyperparameter optimization is performed using randomized search (Bergstra and Bengio, 2012) over 5-fold cross validation over the training data.

3 Experiments

Dataset. We experiment with the ASR data released for the 2012 and 2013 editions of the IWSLT evaluation campaign (Federico et al., 2012; Cettolo et al., 2013) respectively consisting of 11 and 28 English TED talks. The 2012 test set, which has a total speech duration of around 1h45min, contains $1,118$ reference sentences and $18,613$ running words. The 2013 test set has a total duration of around 3h55min, it contains $2,238$ references and $41,545$ running words. In our experiments, we always use $1,118$ utterances for training and $1,120$ utterances for testing. To this aim, the (larger) IWSLT 2013 test set is randomly sampled three times in training and test sets of such dimensions. The use of two datasets is motivated by the objective of measuring variations in the classification performance of our quality estimators under different conditions: *i)* homogeneous training and test data from the same edition of the campaign, and *ii)* heterogeneous training and test data from different editions of the campaign. All utterances have been transcribed with the systems described in (Falavigna et al., 2012; Falavigna et al., 2013).

Evaluation Metric. As mentioned in §2, we need to assess classification performance with potentially imbalanced data distributions. It has been shown that a number of evaluation metrics for binary classification (*e.g.* accuracy, F-measure, etc.) are biased and not suitable for imbalanced data (Powers, 2011; Zamani et al., 2015). For this reason, we use the balanced accuracy (BA – the average of true positive rate and true negative rate), which equally rewards the correct classification on both classes (Brodersen et al., 2010).

Baseline and Terms of Comparison. The simplest baseline to compare with is a system that always predicts the most frequent class in the training data, which would result in a 50% BA score. Furthermore, we assess the potential of our binary quality estimators against two terms of comparison. The first one is an "*oracle*" obtained by selecting the best label among the output of multiple models. Such oracle is an informed selector able to correctly classify each instance if at least one of the models returns the right class. Significant differences between the performance achieved by the single models and the oracle would indicate some degree of complementarity between the different learning strategies/algorithms. Close results obtained with the stacking method would evidence its capability to leverage such complementarity. The second term of comparison is a *full-fledged quality estimator* that exploits glass-box features as a complement to the black-box ones. Performance differences between the black-box models and the full-fledged estimator will give an idea of the potential of each method both in the most interesting, but less favourable condition (*i.e.* when the ASR system used to transcribe the signal is unknown), and in the most favourable condition when confidence information is also accessible.

Results and Discussion. We evaluate our approach in two experimental setups, characterized by different distributions of positive/negative instances. These are obtained by setting the threshold τ to 0.05 and 0.4. In both settings, the minority class contains around 20% of the data in the majority class. Tables 1 and 2 show the results obtained by: *i)* models trained and evaluated on data either from the same (2013-2013) or different editions of IWSLT (2012-2013), and *ii)* models trained using either ALL the features (*i.e.* glass-box and black-box) or only the black-box ones (BB). For the sake of brevity, only the performance of the best classification algorithms is provided, together with the stacking and oracle results. In order to eyeball the significance of the difference in mean values, for each result we also report the standard deviation.

The analysis of the results yields several find-

Table 1: Balanced accuracy (BA) obtained by different methods when $\tau = 0.05$

Train – Test	Features	Classification via regression	Standard classification	Stacking	Oracle
2013 – 2013	BB	SVR: 55.91 ± 3.06	XTC: 66.78 ± 0.18	**76.71 ± 2.23**	85.12 ± 1.69
2013 – 2013	ALL	SVR: 62.76 ± 1.46	SVC: 77.31 ± 1.33	**86.33 ± 1.93**	90.87 ± 1.03
2012 – 2013	BB	XTR: 50.00 ± 0.0	SVC: 62.34 ± 1.97	**75.90 ± 2.54**	85.63 ± 1.11
2012 – 2013	ALL	XTR: 61.72 ± 0.38	MaxEnt: 75.82 ± 0.85	**88.40 ± 0.74**	90.36 ± 0.61

Table 2: Balanced accuracy (BA) obtained by different methods when $\tau = 0.4$

Train – Test	Features	Classification via regression	Standard classification	Stacking	Oracle
2013 – 2013	BB	SVR: 68.49 ± 1.03	SVC: 72.29 ± 0.58	**78.47 ± 3.77**	86.65 ± 0.31
2013 – 2013	ALL	XTR: 76.63 ± 0.54	SVC: 80.43 ± 0.19	**88.06 ± 1.98**	89.11 ± 1.45
2012 – 2013	BB	XTR: 54.67 ± 1.21	SVC: 62.60 ± 2.06	**76.17 ± 2.78**	81.85 ± 1.74
2012 – 2013	ALL	SVR: 69.19 ± 0.62	MaxEnt: 80.02 ± 0.54	**87.34 ± 1.49**	90.80 ± 0.38

ings, relevant from the application-oriented perspective that motivated our research. First, in all the testing conditions our best binary classifiers significantly outperform the majority class baseline (50% BA). Top results with homogeneous data (2013-2013) are up to *78.47%* when only black-box features are available, and *88.40%* when all the features are combined. Not surprisingly, the scores achieved by classifiers trained only with BB features are lower than those achieved by models that can leverage ALL the features. Nevertheless, the positive results achieved by the BB features indicate their potential to cope with the difficult condition in which the inner workings of the ASR system are not known.

As regards the different learning strategies, a visible trend can be observed: standard classification significantly outperforms classification via regression in all cases. This indicates that it substantially benefits from the instance weighting mechanism described in §2, and from the fact that model selection can be performed by maximizing BA (the same metric used to evaluate the system), which cannot be used by the classification via regression strategy. Regarding the algorithmic aspect, the analysis of the results does not lead to definite conclusions. Indeed, none of the tested algorithms seems to consistently prevail across the different testing conditions and, especially for classification via regression, the best score is often not significantly better than the others. Looking at the oracle, its high BA suggests a possible complementarity between the different strategies/algorithms, and large room for improvement over the base estimators. Such complementarity is successfully exploited by the stacking method, which drastically reduces the gap in all cases.

All the learning strategies suffer from evaluation settings where the data distribution is heterogeneous (2012-2013). Although the oracle results do not show large differences when moving from the 2013-2013 to the 2012-2013 setting, almost all the results show consistent performance drops in the latter, more challenging setting. Nonetheless, the BA achieved by the stacking method is rather high, and always above *75%*.

4 Conclusions

We investigated the problem of assigning informative and unambiguous binary quality labels (good/bad) to automatically transcribed utterances. Aiming at an application-oriented approach, we developed a *reference-free* and *confidence-independent* method, which has been evaluated in different settings. Our experiments on English TED talks' transcriptions from the IWSLT campaign show that our best *stacking* models can successfully combine the complementarity of different strategies. With a balanced accuracy ranging from *86.33%* to *88.40%*, the full-fledged classifiers that combine black-box and glass-box (*i.e.* confidence-based) features bring the problem close to its solution. With results in the range *75.90%-78.47%*, our reference-free and confidence-independent models provide a reliable solution to meet the demand of cost-effective methods to estimate the quality of the output of unknown ASR systems.

References

J. Bergstra and Y. Bengio. 2012. Random Search for Hyper-parameter Optimization. *J. Mach. Learn. Res.*, 13(1):281–305.

K. H. Brodersen, C. S. Ong, K. Enno Stephan, and J. M. Buhmann. 2010. The Balanced Accuracy and Its Posterior Distribution. In *Proceedings of the 20th International Conference on Pattern Recognition*, ICPR '10, pages 3121–3124, Istanbul, Turkey.

José G. C. de Souza, Marco Turchi, and Matteo Negri. 2014. Machine Translation Quality Estimation Across Domains. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING 2014): Technical Papers*, pages 409–420, Dublin, Ireland, August.

José G. C. de Souza, Hamed Zamani, Matteo Negri, Marco Turchi, and Daniele Falavigna. 2015. Multitask learning for adaptive quality estimation of automatically transcribed utterances. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 714–724, Denver, Colorado.

M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Federico. 2013. Report on the 10th IWSLT Evaluation Campaign. In *Proceedings of the International Workshop for Spoken Language Translation*, IWSLT '13, Heidelberg, Germany.

C. Cortes and V. Vapnik. 1995. Support-Vector Networks. *Mach. Learn.*, 20(3):273–297.

I. Csiszár. 1996. Maxent, Mathematics, and Information Theory. In *Proceedings of the Fifteenth International Workshop on Maximum Entropy and Bayesian Methods*, pages 35–50, Sante Fe, New Mexico, USA.

G. Evermann and P. C. Woodland. 2000. Large Vocabulary Decoding and Confidence Estimation Using Word Posterior Probabilities. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP '00, pages 2366–2369, Istanbul, Turkey.

D. Falavigna, G. Gretter, F. Brugnara, and D. Giuliani. 2012. FBK @ IWSLT 2012 - ASR Track. In *Proc. of the International Workshop on Spoken Language Translation*, Hong Kong, HK.

D. Falavigna, R. Gretter, F. Brugnara, D. Giuliani, and R. Serizel. 2013. FBK@IWSLT 2013 - ASR Tracks. In *Proc. of IWSLT*, Heidelberg, Germany.

M. Federico, M. Cettolo, L. Bentivogli, M. Paul, and S. Stüker. 2012. Overview of the IWSLT 2012 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, Hong Kong, HK.

P. Geurts, D. Ernst, and L. Wehenkel. 2006. Extremely Randomized Trees. *Mach. Learn.*, 63(1):3–42.

Shahab Jalalvand, Matteo Negri, Falavigna Daniele, and Marco Turchi. 2015. Driving rover with segment-based asr quality estimation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1095–1105, Beijing, China.

A. Mammone, M. Turchi, and N. Cristianini. 2009. Support vector machines. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(3):283–289.

L. Mangu. 2000. *Finding Consensus in Speech Recognition*. Ph.D. thesis, John Hopkins University.

Yashar Mehdad, Matteo Negri, and Marcello Federico. 2012. Match without a Referee: Evaluating MT Adequacy without Reference Translations. In *Proceedings of the Machine Translation Workshop (WMT2012)*, pages 171–180.

N. Meinshausen and P. Bhlmann. 2010. Stability Selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473.

M. Negri, M. Turchi, and J. G. C. de Souza. 2014. Quality Estimation for Automatic Speech Recognition. In *Proceedings of the 25th International Conference on Computational Linguistics*, COLING '14, Dublin, Ireland.

D. M. W. Powers. 2011. Evaluation: From Precision, Recall and F-measure to ROC, Informedness, Markedness & Correlation. *J. Mach. Learn. Tech.*, 2(1):37–63.

M. S. Seigel and P. C. Woodland. 2014. Detecting Deletions in ASR Output. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP '14, pages 2321–2325, Florence, Italy.

M. S. Seigel. 2013. *Confidence Estimation for Automatic Speech Recognition Hypotheses*. Ph.D. thesis, Cambridge University.

Lucia Specia, Nicola Cancedda, Marc Dymetman, Marco Turchi, and Nello Cristianini. 2009. Estimating the Sentence-Level Quality of Machine Translation Systems. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT'09)*, pages 28–35, Barcelona, Spain.

Y. C. Tam, Y. Lei, J. Zheng, and W. Wang. 2014. ASR Error Detection Using Recurrent Neural Network Language Model and Complementary ASR. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP '14, pages 2331–2335, Florence, Italy.

Marco Turchi, Antonios Anastasopoulos, José G. C. de Souza, and Matteo Negri. 2014. Adaptive Quality Estimation for Machine Translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 710–720, Baltimore, Maryland.

K. Veropoulos, C. Campbell, and N. Cristianini. 1999. Controlling the Sensitivity of Support Vector Machines. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, IJCAI '99, pages 55–60, Stockholm, Sweden.

F. Wessel, K. Macherey, and R. Schlüter. 1998. Using Word Posterior Probabilities as Confidence Measures. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP '98, pages 225–228, Seattle, Washington.

D. H. Wolpert. 1992. Stacked Generalization. *Neural Networks*, 5(2):241–259.

H. Xu, D. Povey, L. Mangu, and J. Zhu. 2010. An Improved Consensus-Like method for Minimum Bayes Risk Decoding and Lattice Combination. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, ICASSP '10, Dallas, Texas, USA.

Hamed Zamani, Pooya Moradi, and Azadeh Shakery. 2015. Adaptive user engagement evaluation via multi-task learning. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 1011–1014, Santiago, Chile.