How to evaluate ASR errors impact on NER?

Mohamed Ameur Ben Jannet[1,2,3,4], *Olivier Galibert*[2], *Martine Adda-Decker*[3], *Sophie Rosset*[4]

[1] Université Paris-Sud
[2]LNE
[3]LPP–CNRS UMR 7018, Université Sorbonne Nouvelle
[4]LIMSI–CNRS UPR 3251

{first.last}@limsi.fr, {first.last}@lne.fr

Abstract

The standard metric to evaluate automatic speech recognition (ASR) systems is the word error rate (WER). WER has proven very useful in stand-alone ASR systems. Nowadays, Speech recognition is one of the most widely used components in spoken language processing applications. Its hypotheses are a valuable source of features for downstream modules which try to reach the semantics of the message. Therefore, ASR systems are often embedded in complex natural language processing systems to perform tasks like speech translation, man-machine dialogue, or information retrieval from speech. This exacerbates the need for the speech processing community to design a new evaluation methodology to estimate the quality of automatic transcriptions and the impact of transcription errors within their larger applicative context.

In this paper we consider the case of ASR systems which output is used for named entities recognition (NER). We start by discussing the strengths and limits of the standard WER and NE-WER metrics in that context. Then we took advantage of the French ETAPE and QUAERO evaluation campaigns results which included named entities recognition from ASR output to analyse the impact of WER on NER performances. We show that ASR errors may have different impacts, reducing recall or precision, which are not directly deducible from WER changes. We then introduce a new metric, Automatic Transcription Evaluation for Named Entities (ATENE) to evaluate the ASR in the context of named entity recognition. Our proposed metric makes use of a probabilistic model to estimate the risk of ASR errors inducing downstream errors. ATENE includes two elementary measures, $ATENE_{DS}$, relative to deletions and substitutions, to evaluate the variation of recall caused by ASR errors and $ATENE_{I}$, relative to insertions, to evaluate the variation of precision. We compare our new metric to the standard one on the ETAPE and QUAERO evaluation outputs. ATENE achieves a higher correlation than WER and NE-WER showing that it can be used as alternative to standard measures to select the best ASR system for NER.

Index Terms: speech recognition, ATENE, named entity recognition, metric

1. Introduction

Tremendous progress has been observed during the last decades, for example for open-vocabulary and continuous speech recognition (see [1] or [2]) or robustness against speakers' variation and noisy environment (see [3] or [4]). The systems have become performant enough to be embedded in applications such as speech-to-speech translation, spoken information retrieval or spoken language dialog systems. However the transcription process still entails errors, mainly due to challenging acoustic conditions, out-of-vocabulary words or language ambiguities. The resulting errors are of varying importance for the overall application in which the ASR system is embedded. Our objective is to provide an alternative metric which measures the fitness of the ASR output to the overall task better than the generic ASR-centered metric WER.

In this paper we consider the case of ASR systems which output is used for named entities recognition (NER). Through previous works, we first discuss in Section 2 the adequacy of existing metrics to evaluate ASR transcriptions for NER, validating that WER can be improved upon for this aim.

In section 4 we describe the data used in this work along with an analysis of the impact of WER on NER performance.

We then introduce, in Section 5, a new metric, Automatic Transcription Evaluation for Named Entities (ATENE), to evaluate ASR in the context of named entity recognition. We compare our new metric to the standard ones in Section 6 and conclude this paper in Section 7.

2. Related Work

The main ASR metric is the WER, which counts the errors in the transcription and normalizes it by the size of the reference. The different errors are substitutions, deletions and insertions of word, determined by a Levenstein alignment [5] of reference and hypothesis transcriptions. The WER is thus an error-enumeration based metric which considers every error as equally important. One wonders whether this approach is the most appropriate to evaluate, and choose ASR systems given one specific applications. To answer that question, the correlation between WER and the performance obtained by the overall application was measured. For example, in the context of webcast archives, the influence of WER on the usability and usefulness of the archives was investigated in [6]. Their results showed that speech recognition accuracy linearly influenced users' performance in the task of quiz answering. Other studies focused on performance of NLP system working on such outputs ([7] in the context of an information retrieval task, [8] in the context of speech translation and [9] in the context of spoken language understanding). They have shown that the WER is not always well correlated with the application performance. Some alternatives metrics to the WER have been proposed. In [10], it was proposed to measure the loss of information caused by ASR errors. The Relative Information Loss (RIL), is a stochastic based measure which uses the difference of entropy between the hypothesis words as such and in the context of the reference:

$$RIL = \frac{H(Y|X)}{H(Y)} \tag{1}$$

Where X is the reference, Y the hypothesis and H the normal entropy estimation on a word vector:

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{2}$$

$$H(Y|X) = -\sum_{i,j} P(x_i|y_j) \log P(x_i|y_j) \tag{3}$$

Probability in (2) and (3) can be estimated from the relative frequencies through an alignment between reference and hypothesis. RIL has two main problems: it considers systematic words substitutions as correct, and its implementation is very complex [11]. Word Information Lost (WIL) has been introduced in [12] as an approximation of RIL. For high error rates [12] and [13] found that RIL and WIL can be appropriate. The evaluation of ASR performances for information retrieval from spoken documents has drawn much attention, especially in the context of the TREC evaluation. In [7] it was shown that there is a high correlation between WER and the performance in information retrieval, but it was noted that ranking ASR systems based on the WER and ranking them based on the performances obtained on information retrieval gave a different order, the best performance not always being achieved on the best ASR output. That shows that while the WER can help in predicting the performance impact of the ASR errors it is not always sufficient to select the best ASR transcription for the retrieval task, especially when the ASR systems show rather close WER scores. All that suggests that the performance in information retrieval depends not only on the amount of errors but also on their types. In [14] the authors also compare with the Named Entity Word Error Rate (NE-WER), which consists of a normal WER restricted to the words of the reference present in a named entity (NE). The correlation with the IR results was higher, but the system rankings were not changed, not making the metric significantly better for system selection. One possible cause is that NE-WER ignores inserted or substituted words outside of NE which cause false alarms in the downstream IR.

The previous work suggests that spoken language processing applications, dependent on ASR performance, would benefit for an ASR system optimized for the applicative case, and that WER is not the best metric to achieve that aim.

3. The Named Entities Recognition Task

Since its creation in the MUC conferences [15], the Named Entities Recognition task has become a critical step in numerous language processing applications [16]. The task consists in detecting, classifying and decomposing all mentions of named entities which are, in an intuitive approximation, the objects of the real world the discourse is referring to. Numerous annotation schemes with varying complexity and coverage exists. For this study we concerned ourselves with the Quaero [17, 18] scheme. It has the advantage to propose a structural complexity and a coverage higher than most other schemes, so that a metric validated on it will be robust to other task variants.

The taxonomy involves seven classical named entities classes: person, locations, organizations, functions, products, temporal expressions and amounts. Annotations are on two levels, the first level for a full entity classification and the second for a decomposition of the entity contents in different typed slots. The taxonomy is hierarchical with a number of subtypes, and the annotation recursive, entities may be included in other entities.

That schema has been used in two evaluations, in 2010 within Quaero [19] and in 2014 with the open campaign ETAPE [20]. ETER (Entities Tree Error Rate) [21] is the current metric to evaluate the task in both manual and automatic transcription conditions:

$$ETER = \frac{I + D + \sum_{(e_r,e_h)} E(e_r, e_h)}{N} \tag{4}$$

where I and D represent the number of inserted and deleted entities determined through an alignment of reference and hypotheses annotations. $E(e_r, e_h)$ is the sum of classification and decomposition errors for matched entities betwen the reference and the hypothesis. And N is the total number of entities in the reference. That metric is similar to the Slot Error Rate [22] with a more elaborate classification/decomposition error estimation. It is essentially an error-enumeration metric, making it quite close in its fundamentals to WER. An interesting property that the formula shows is that the metric can be decomposed in its individual parts, insertions (I), deletions (D) and substitutions (E). We will use it in our correlation evaluations Section 6.2.

4. Data Description and interpretation of ASR errors impact

4.1. Data description

For our experiments we used the QUAERO [23] and ETAPE [24] data sets. These two corpora are fully annotated in named entities according to the [17] guidelines. They also both have been used for an evaluation of NER on automatically transcribed speech, in the QUAERO [19] and ETAPE [20] benchmarks. As a result multiple automatic transcriptions, and multiple NER runs from very different systems are available in these data sets. Table 1 provides some statistics for both corpus in terms of number of words and named entities.

	ETAPE		QUAERO	
	Train	Test	Train	Test
Words	335 387	115 803	1 251 586	97 871
Ents.	19 270	5 933	113 885	5 523

Table 1: Statistics on words and entities of the ETAPE and QUAERO test corpora.

The ETAPE test data set was composed of 15 files, each one instance of a broadcast news or broadcast conversations program fully transcribed by five different ASR systems plus a (medium quality) rover, built by a vote between the different hypothesis. All were annotated by seven different NER systems in extended named entities. Two of the NER systems used a symbolic approach and the others were based on stochastic or hybrid approaches. All the data, including system outputs, is in the process of being added to the ELRA catalogue for wide availability. In the QUAERO case the test data set was composed of 18 files, each one instance of a broadcast news or conversations program and was transcribed by three different ASR systems, and annotated by three NER systems, one symbolic, one purely stochastic and one hybrid. The training and test data are available through ELDA, with all ASR outputs and one NER system output.

4.2. Interpretation of the ASR errors impact

Figure 1 (left) summarizes the ETAPE benchmark results. On this figure we can observe that the rover enables the best performances for NER even if with a WER of 28% (worse than 3 of the ASR systems). We can also see that the ASR-2 transcriptions (25% WER) end up in better results then ASR-1 (22% WER) for four NER systems (NER-3, 4, 5 and 7). The ASR-4 transcription (30% WER) still allow comparable performances to those obtained on ASR-3 (26% WER).

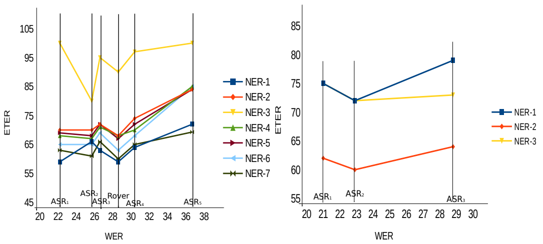

Figure 1: Cross-Recognizer Results: ETER (Entities Tree Error Rate) vs. WER (Word Error Rate), ETAPE (left) and QUAERO (right) benchmarks results.

Figure 1 (right) shows the QUAERO benchmark results. The best performance in NER was obtained on the ASR output ranked second in terms of WER.

To better understand the impact of ASR errors (insertions, deletions, substitutions) on NER behavior, we want to determine if these ASR errors induce the same kind of errors in NER systems applied downstream. This can be done by measuring the increase of each kind of NER error (deletions, insertions and substitution of named entities) caused by ASR. PEI (Percentage of Error Increase) is calculated as follows:

$$PEI(e) = 100 * \frac{Nb_H(e) - Nb_R(e)}{Nb_R(e)}$$

Where e is the kind of NER error which can be deletion, insertion or substitution. $Nb_H(e)$ and $Nb_R(e)$ are respectivly the number of NER errors (of type e) on automatic transcription and on the reference (manual transcription).

Figure 2 shows the PEI obtained by the three NER systems that participated in the QUAERO evaluation.

We can see that for all NER systems, the ASR-1 transcription (in blue) lead to an increase of insertions errors but in the same time it causes less deletions. The ASR-2 transcription (in red) leads to the opposite behavior, causing less insertions but more deletions. This suggests that a different WER (or ASR system performance) does not simply lead to a difference in overall NER system performance but also and especially to a difference in the types of errors.

Moreover, the ETAPE and QUAERO benchmarks results show that an increase of WER does not necessary induce the rise of ETER for the NER systems. This observation strengthens our claim for an alternative metric taking into account the applicative context and the impact of transcriptions errors on NER systems.

5. Proposition

5.1. Principle

To evaluate the quality of the ASR output in the NER context, we need to quantify the impact of the errors on the detection. Generally speaking, to classify the entities, the NER systems rely on multiple levels of contextual features such as words, parts of speech, or lexical information such as proper names, dictionaries or gazetteers. The ASR errors change these features and thus influence the NER decisions.

But examining the errors is not sufficient. The impact of an error depends on its nature, but also on the context in which it occurs. It is important to take into account the whole contextual information when measuring the impact of an ASR error.

Rather than directly comparing reference and hypothesis transcriptions, we propose to measure how harder it became to identify entities given the differences between hypothesis and reference by comparing an estimated likelihood of presence of entities. The estimation is obtained through a statistical classifier which uses basic features (words, prefixes and suffixes) common to symbolic and stochastic approaches in NER. The use of simple features provides system and approach independence and avoids the considerable cost of development of a state-of-the-art NER system.

Looking only at the top-level annotation, we can label words depending on whether it is present in an entity and its type (person, location...). We intend to measure the difficulty of distinguishing the correct answer by computing the *margin*, which is the difference in probability between the reference label $P(\hat{Y})$ and the probability of the most likely incorrect label $\max_{Y \neq \hat{Y}} P(Y)$.

$$M(X) = P(\hat{Y}|X) - \max_{Y \neq \hat{Y}}(P(Y|X)) \qquad (5)$$

where X is the vector of features (words, prefixes, suffixes) at a given position in the text. In order to estimate the change in difficulty, we compute the difference between the margin at a given position in the ASR output and the margin at the same position in the reference transcription. A negative ΔM means that errors make the task more difficult, positive less.

$$\Delta M(X_A, X_R) = Marg(X_A) - Marg(X_R) \qquad (6)$$

Where $X_A \; and \; X_R$ are vectors of features extracted from the same position in ASR transcripts and in the reference.

5.2. Entity projection

Computing ΔM requires being able to align positions between the reference manual transcription and the ASR output. We do that by extending the approach defined in [25]. The first step consists in a forced acoustic alignment of the reference text on the signal, providing a temporal position for every word. We use these positions with a margin to find possible spans for the entities in the ASR output. Finally the choice between these spans is done by selecting the word string the closest to the reference after conversion to phonetic strings through a pronunciation dictionary. The procedure gives us positional associations between reference and hypothesis for the words at the start and the end of every entity.

5.3. $ATENE_{DS}$

Two kinds of NER errors can happen where reference entities are present: deletions and substitutions. The first case is a consequence of the no-annotation probability going up, the other of

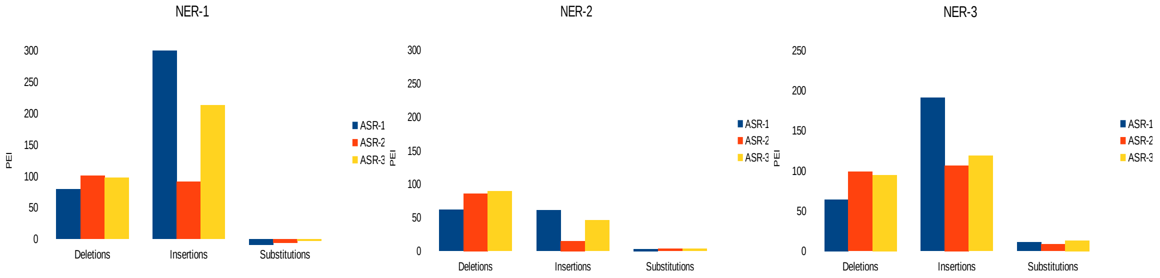

Figure 2: Percentage of NER errors (Deletions, Insertions, Substitutions) increase according to ASR systems transcriptions, QUAERO benchmark results.

a wrong label probability going up. Both cases do not need to be distinguished.

For the NER task, correctly detecting an entity consists, in a large part, of detecting its start and its end. Since the alignment procedure gave us the positions in the hypothesis and the reference for the starts and ends, we can apply the ΔM computation there. We use two separate models, one for starts where the first word of an entity is labelled with the entity type and the others with a null label, and one similar for the end of entities. We can then compute the mean of the probability margin differences:

$$ATENE_{DS} = \frac{\sum_{i=1}^{N} \Delta M(Start_i) + \Delta M(End_i)}{2N} \tag{7}$$

Where N is the total number of entities in reference. We expect a large negative value to be correlated to a high rate of deletions and substitutions, e.g. a low recall.

5.4. $ATENE_I$

In complement to the deletion and substitution NER errors we need to estimate the chances of getting extra insertions induced by the ASR errors. The projection process has given us a decomposition of the ASR hypothesis in segments that are alternatively inside and outside an entity. It is thus possible to collate all hypothesis and reference word segments that are present outside of any entity and put them in 1-1 relation. What we do not have is a 1-1 relation between words, if only because the size of the segments changes between reference and hypothesis with the ASR insertions and deletions.

Statistics on our development data has shown that, for a given inter-entity span, insertion errors count is 0 or 1 more than 90% of the time. So for each segment we estimate the risk of having at least one insertion error by finding the lowest correct detection margin for all words in the span. We use a model where the word labels are either *out-of-entity* or the entity type.

$$M_O(S) = \min\left(P(O|X_i) - \max_{Y!=O} P(Y|X_i)\right) \tag{8}$$

Where S is an inter-entities segment, X_i is the vector of features at $word_i$ of segment S, and O is the out-of-entity label. Since we aim at detecting errors, we decided to take the margin into account only when negative, e.g. when an error seems possible. Otherwise the margin is set to 1:

$$M'_O(S) = \begin{cases} M_O(S) & M_O(S) < 0 \\ 1 & otherwise \end{cases} \tag{9}$$

Following the $ATENE_{DS}$ stucture we then compute the margin difference between reference and hypothesis, and compute the mean of all segments:

$$\Delta M_O(S) = M'_O(S_A) - M'_O(S_R) \tag{10}$$

Where $S_A \; and \; S_R$ are inter-entity segments matched between ASR transcript and reference.

$$ATENE_I = \frac{\sum_{i=1}^{|S|} \Delta M_O(S_i)}{|S|} \tag{11}$$

We expect a large negative value to be correlated with a high rate of insertion, e.g. a low precision.

5.5. Final score

In order to compute a unique score to the whole ASR transcriptions for NER, the global score $ATENE$ is the mean of the two $ATENE_{DS}$ and $ATENE_I$ scores. Both $ATENE_{DS}$ and $ATENE_I$ compute a number of measurements valued between -1 and 1 and take their mean. The measurements count is equal or very close to the number of entities in the reference in both cases. So there is a fair chance that the two halves behave compatibly and the simple mean is the correct method.

$$ATENE = -100\frac{ATENE_{DS} + ATENE_I}{2} \tag{12}$$

The final -100 multiplier is added for the readability of the values, to reach a behaviour similar to an error rate, e.g. the higher the final value the worse the score is.

5.6. Classifier

This methodology relies on statistical models predicting the presence of named entities in a given context. The models need to be applied at specific positions in the text and provide probabilities. In addition we have no specific hypothesis on the structure of the stochastic landscape of the features. Maximum entropy models are a natural fit for such needs. The models were trained on the QUAERO and ETAPE NER training corpora using the Wapiti toolbox [26]. Standard features were used:

- Words and bi-grams of words in a [-2,+2] window of around the target word
- Prefixes and suffixes of words in a [-1,+1] window of around the target word

These features are a subset, and in fact the core, of what both symbolic and stochastic systems look at. Capitalization is also a feature NER systems leverage, but not all ASR systems provide one. So we created variants of the models, one ignoring capitalization, one taking it into account. In our data one of the ETAPE and all three QUAERO ASR outputs are capitalized.

Metric (w.r.t. $ETER$)	NER-1	NER-2	NER-3	NER-4	NER-5	NER-6	NER-7	mean
WER	0.57	0.52	0.29	0.51	0.48	0.55	0.53	0.49
NE-WER	0.79	0.68	0.31	0.60	0.67	0.64	0.59	0.61
ATENE	0.74	0.82	0.45	0.66	0.77	0.78	0.73	0.71
Metric (w.r.t $ETER_{DS}$)	NER-1	NER-2	NER-3	NER-4	NER-5	NER-6	NER-7	mean
WER	0.47	0.66	0.70	0.59	0.76	0.70	0.62	0.64
NE-WER	0.69	0.69	0.43	0.47	0.67	0.51	0.67	0.59
$ATENE_{DS}$	0.63	0.75	0.56	0.56	0.77	0.70	0.73	0.67
Metric (w.r.t $ETER_I$)	NER-1	NER-2	NER-3	NER-4	NER-5	NER-6	NER-7	mean
WER	0.34	0.38	-0.16	0.27	0.21	0.34	0.35	0.25
NE-WER	0.47	0.48	0.22	0.56	0.46	0.53	0.43	0.45
$ATENE_I$	0.75	0.76	0.28	0.75	0.76	0.85	0.74	0.70

Table 2: Mean Spearman correlation between $ETER$/$ETER_{DS}$/$ETER_I$ and the score given by WER, NE-WER and $ATENE$/$ATENE_{DS}$/$ATENE_I$ on the ETAPE challenge outputs.

6. Experiments and Results

6.1. Evaluation methodology

Comparing ASR systems based on the quality of their transcription for NER implies to rank them according to the ETER results obtained on their output. Therefore, we consider the ranking of the ASR transcripts keyed on a given NER system performance as ground truth. We can then measure the correlation between this reference rank and the ranks obtained through WER, NE-WER and ATENE to evaluate which measure predicts best the ASR output quality for NER task.

The most popular methods for calculating the correlation of ranked results lists are Kendall's τ and Spearman's ρ [27]. Spearmans correlation is reflecting the degree of concordances and discordances on the rank scale, whereas Kendalls τ correlation coefficient reflects only the numbers of concordances and discordances regardless of their degree [28]. We propose to use Spearman ρ because it handles the case where tied ranks are present. We noticed though that using Kendall's tau did not change the relative results, only the absolute values.

Multiple measurement points make for more robust results. Having only six ASR outputs for ETAPE and three for QUAERO gives a relatively high intrinsic imprecision in the ranking correlation. To reduce it we decided to compute the mean of the correlations for every show comprising a test set independently. Similarly we took the mean of the correlation obtained on each NER system to compute a global score.

6.2. Results

Tables 2 shows that our proposed metric ATENE correlates better than the WER and NE-WER with the results of most NER systems (NER-1 is the only exception, and not by much). The global mean correlation is also the best for our metric. It's interesting that the NER-5 system is a pure symbolic one. The mean correlation of 0.77 on that system shows that the metric's stochastic roots do not bias against symbolic systems.

To refine our understanding we also measured the correlations with the different elements of the metric. The ETER metric computation can be split in two parts, one with deletions and substitutions ($ETER_{DS}$), the other with insertions only ($ETER_I$). That allows us to compute the correlations with the two halves of the ATENE metric.

The additional Tables 2 confirm the quality of the $ATENE_{DS}$ metric to estimate the risks of deletion and substitution. Still we can see that the NE-WER is a decent, if not as good, estimator of that risk. That seems reasonable since it restricts measuring errors to the entity spans. They also show that $ATENE_I$ is a good estimator of the insertion risk. WER-NE is nowhere near as bad as expected since that metric does not measure anything on the words where insertions can happen.

In every case we also see that the WER metric is less performant than the specialized ones, validating our primary hypothesis that the WER is insufficient for a multi-module application.

The ETAPE data has been used during the design of the metric and has in part influenced its design, making it more of a development set than a real test set. For a robust validation we needed to compute the same correlations on a not-previously seen corpus. We kept the QUAERO corpus for that.

Metric (w.r.t $ETER$)	NER-1	NER-2	NER-3	mean
WER	0.38	0.36	0.11	0.28
NE-WER	0.19	0.16	-0.02	0.11
$ATENE$	0.44	0.50	0.55	0.50
Metric (w.r.t $ETER_{DS}$)	NER-1	NER-2	NER-3	mean
WER	0.72	0.52	0.77	0.67
NE-WER	0.80	0.58	0.77	0.72
$ATENE_{DS}$	0.66	0.55	0.63	0.62
Metric (w.r.t $ETER_I$)	NER-1	NER-2	NER-3	mean
WER	0.08	-0.13	-0.36	-0.13
NE-WER	-0.19	-0.41	-0.52	-0.37
$ATENE_I$	0.50	0.55	0.22	0.42

Table 3: Mean Spearman correlation between $ETER$/$ETER_{DS}$/$ETER_I$ and the score given by WER, NE-WER and $ATENE$/$ATENE_{DS}$/$ATENE_I$ on QUAERO challenge outputs.

Tables 3 shows that ATENE still has the best correlations with the global performance of NER systems on QUAERO data. Interestingly WER and NE-WER work rather well on the deletion/substitution subproblem and achieve even better results than our elementary measure $ATENE_{DS}$. This suggests that WER and NE-WER are good options to evaluate the impact of ASR errors on NER recall. But they're completely unusable when it comes to predict the insertions.

7. Conclusion

This paper addressed the issue of evaluating the quality of ASR transcriptions in a complex NLP task combining ASR with

NER. Unfortunately, standard metrics, such as WER and NE-WER show a relatively low correlation between ASR and NER performances using Spearman's rho rank correlation. This result is not so surprising as the WER metric was not designed to care about post-ASR processing tasks when evaluating ASR transcripts. With respect to the NE-WER metric, a major weakness consists in not properly taking into account the risk of false alarm errors when evaluating ASR transcripts for NER.

To overcome this limitation and to better account for the applicative task context in the ASR evaluation, ATENE is measuring the risk of errors in downstream modules as induced by the ASR mistakes. The different kinds of error (deletions, substitutions and insertions) that transcription entail in downstream NER systems are taken into account. The merits of ATENE were tested by comparing it to WER and NE-WER on ETAPE and QUAERO benchmark data. ATENE achieves a higher correlation with the NER results, showing the added value of this new metric. Future work includes the optimization of the ASR system settings with respect to our metric.

8. Acknowledgements

This work was funded by the ANR VERA project (adVanced ERror Analysis for speech recognition) funded by ANR Blanc (ANR 12 BS02 006 04) and the CIFRE grant No 2012/0771. This work was also supported by the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083).

9. References

[1] M. Gerosa and M. Federico, "Coping with out-of-vocabulary words: open versus huge vocabulary asr," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*. IEEE, 2009, pp. 4313–4316.

[2] C. Parada, M. Dredze, A. Sethy, and A. Rastrow, "Learning sub-word units for open vocabulary speech recognition," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 712–721.

[3] A. L. Maas, Q. V. Le, T. M. O'Neil, O. Vinyals, P. Nguyen, and A. Y. Ng, "Recurrent neural networks for noise reduction in robust asr." in *INTERSPEECH*, 2012.

[4] F. Weninger, M. Wollmer, J. Geiger, B. Schuller, J. F. Gemmeke, A. Hurmalainen, T. Virtanen, and G. Rigoll, "Non-negative matrix factorization for highly noise-robust asr: to enhance or to recognize?" in *ICASSP*. IEEE, 2012, pp. 4681–4684.

[5] V. I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," *Soviet physics doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[6] C. Munteanu, R. Baecker, G. Penn, E. Toms, and D. James, "The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 2006, pp. 493–502.

[7] J. S. Garofolo, C. G. Auzanne, and E. M. Voorhees, "The TREC spoken document retrieval track: A success story." *NIST SPECIAL PUBLICATION SP*, vol. 500, no. 246, pp. 107–130, 2000.

[8] X. He, L. Deng, and A. Acero, "Why word error rate is not a good metric for speech recognizer training for the speech translation task?" in *ICASSP*. IEEE, 2011, pp. 5632–5635.

[9] Y.-Y. Wang, A. Acero, and C. Chelba, "Is word error rate a good indicator for spoken language understanding accuracy," in *Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*. IEEE, 2003, pp. 577–582.

[10] G. A. Miller, "Note on the bias of information estimates," *Information theory in psychology: Problems and methods*, vol. 2, pp. 95–100, 1955.

[11] V. Maier, "Evaluating ril as basis of automatic speech recognition devices and the consequences of using probabilistic string edit distance as input," *Univ. of Sheffield, third year project*, 2002.

[12] A. C. Morris, V. Maier, and P. Green, "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition." in *INTERSPEECH*, 2004.

[13] I. A. McCowan, D. Moore, J. Dines, D. Gatica-Perez, M. F lynn, P. Wellner, and H. Bourlard, "On the use of information retrieval measures for speech recognition evaluation," IDIAP, Tech. Rep., 2004.

[14] J. S. Garofolo, E. M. Voorhees, C. G. Auzanne, V. M. Stanford, and B. A. Lund, "1998 trec-7 spoken document retrieval track overview and results," in *Broadcast News Workshop*, vol. 99, 1999, p. 215.

[15] R. Grishman and B. Sundheim, "Message Understanding Conference - 6: A brief history," in *Proc. of COLING*, 1996, pp. 466–471.

[16] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, and J. M. Gómez-Berbís, "Named entity recognition: Fallacies, challenges and opportunities," *Computer Standards & Interfaces*, vol. 35, no. 5, pp. 482 – 489, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0920548912001080

[17] S. Rosset, C. Grouin, and P. Zweigenbaum, "Entités nommées structurées : guide d'annotation quaero. limsi–cnrs, orsay, france," 2011.

[18] C. Grouin, S. Rosset, P. Zweigenbaum, K. Fort, O. Galibert, and L. Quintard, "Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview," in *Proceedings of the 5th Linguistic Annotation Workshop*. Portland, Oregon, USA: Association for Computational Linguistics, June 2011, pp. 92–100. [Online]. Available: http://www.aclweb.org/anthology/W11-0411

[19] O. Galibert, L. Quintard, S. Rosset, P. Zweigenbaum, C. Nédellec, S. Aubin, L. Gillard, J.-P. Raysz, D. Pois, X. Tannier, L. Deléger, and D. Laurent, "Named and specific entity detection in varied data: The Quaero named entity baseline evaluation," in *Proc of LREC*. Valletta, Malta: ELRA, 2010.

[20] O. Galibert, J. Leixa, G. Adda, K. Choukri, and G. Gravier, "The ETAPE speech processing evaluation," in *Proc of LREC*. Reykjavik, Iceland: ELRA, 2014.

[21] M. A. Ben Jannet, M. Adda-Decker, O. Galibert, J. Kahn, and S. Rosset, "Eter : a new metric for the evaluation of hierarchical named entity recognition," in *Proc of LREC*. Reykjavik, Iceland: ELRA, 2014.

[22] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel, "Performance measures for information extraction," in *Proc. of DARPA Broadcast News Workshop*, 1999, pp. 249–252.

[23] S. Galliano, G. Gravier, and L. Chaubard, "The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts," in *Proc of Interspeech 2009*, 2009.

[24] G. Gravier, G. Adda, N. Paulsson, M. Carré, A. Giraudel, and O. Galibert, "The etape corpus for the evaluation of speech-based tv content processing in the french language," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, N. C. C. Chair), K. Choukri, T. Declerck, M. U. ur Do an, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, Eds. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.

[25] O. Galibert, S. Rosset, C. Grouin, P. Zweigenbaum, and L. Quintard, "Structured and extended named entity evaluation in automatic speech transcriptions," in *Proc of IJCNLP*, Chiang Mai, Thailand, 2011.

[26] T. Lavergne, O. Cappé, and F. Yvon, "Practical Very Large Scale CRFs," in *Proceedings the 48^{th} Annual Meeting of the Association for Computational Linguistics*, 2010, pp. 504–513. [Online]. Available: http://www.aclweb.org/anthology/P10-1052

[27] M. G. Kendall, *Rank correlation methods*. Griffin, 1948.

[28] N. S. Chok, "Pearson's versus spearman's and kendall's correlation coefficients for continuous data," Ph.D. dissertation, University of Pittsburgh, 2010.