Generating Task-Pertinent sorted Error Lists for Speech Recognition

Olivier Galibert [1], Mohamed Ameur Ben Jannet [2], Juliette Kahn [1], Sophie Rosset [2]
[1]LNE, F-78190 Trappes, France
[2]LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
{first.last}@lne.fr, {first.last}@limsi.fr

Abstract

Automatic Speech recognition (ASR) is one of the most widely used components in spoken language processing applications. ASR errors are of varying importance with respect to the application, making error analysis keys to improving speech processing applications. Knowing the most serious errors for the applicative case is critical to build better systems. In the context of Automatic Speech Recognition (ASR) used as a first step towards Named Entity Recognition (NER) in speech, error seriousness is usually determined by their frequency, due to the use of the WER as metric to evaluate the ASR output, despite the emergence of more relevant measures in the literature. We propose to use a different evaluation metric form the literature in order to classify ASR errors according to their seriousness for NER. Our results show that the ASR errors importance is ranked differently depending on the used evaluation metric. A more detailed analysis shows that the estimation of the error impact given by the ATENE metric is more adapted to the NER task than the estimation based only on the most used frequency metric WER.

Keywords: Automatic Speech Recognition, Metrics, Error Analysis

1. Introduction

Automatic Speech recognition (ASR) is one of the most widely used components in spoken language processing applications. Its outputs are a valuable source of features for downstream modules which try to reach the semantics of the message. Despite important progress, these systems still produce errors.

Error analysis is one of the keys leading to better systems. There exist a lot of work studying the errors of ASR systems (Rena Nemoto and Adda-Decker, 2008), using knowledge about them in order to improve ASR systems (Boháč et al., 2012; Dufour and Esteve, 2008) or trying to automatically detect them in ASR output (Ghannay et al., 2015).

ASR errors have been mainly investigated in the framework of comparisons between automatic vs. human decoding of speech (Scharenborg, 2007; Lippmann, 1997). They pointed out that although today the best ASR speech models are quite efficient, they have not yet reached the status of being able to perfectly take into account all the observed acoustic variation. Human listeners are still outperforming them by a factor of 5 to 6 (Vasilescu et al., 2012). The taxonomy of errors pointed out that some words are frequently victims of ASR errors: in particular short, acoustically poor and frequent items lead to local ambiguity (Adda-Decker, 2006). Other work was done studying or classifying errors given the type of words which are involved. For example, (Goryainova et al., 2014) studied ASR errors given the Part of Speech of the associated word. Most of the studies done on ASR error analysis focus on the cause of errors more then on the possible impact that it can have on downstream modules. This studies helped to better understand the origine of errors in order to build more robust ASR systèmes. Despite all progresses, ASR systems still not perfect but their performances allows their use in many application case. In the same time the impact of the residual errors still miss understanded, mainly because we dont know how to measure or how to estimate the seriousness of transcription error for modules using ASR output.

ASR errors seriousness can vary with respect to the application (see for example (Comas and Turmo, 2009) for question-answering on speech, or (Dinarelli and Rosset, 2011) for named entity recognition). We place ourselves in the context of Automatic Speech Recognition (ASR) and Named Entity Recognition (NER) combined for a task of NER in speech.

Various evaluation metrics for ASR outputs can be found in the literature. Our hypothesis is that an evaluation metric, besides giving a performance score, is able to provide information about the individual errors produced by an ASR system. We expect a metric to be able to give information about the seriousness of the errors given a task. Thus we are interested by generating a ranking of ASR errors according to different evaluation metrics in order to understand which metric allow a better identification of the most serious ASR errors for NER.

In the following section, we present the ASR evaluation metrics and discuss them in relation with our objective; in Section 3., we present our contribution and in Section 4. the experiments are described along with a discussion.

2. ASR evaluation metrics

A reasonable way of listing ASR errors and estimating their seriousness is to build that based on ASR evaluation metric. The most widely used metric is the word error rate. That metric counts the errors in the transcription and normalizes it by the size of the reference. The different errors are substitutions, deletions and insertions of words, determined by a Levenstein alignment (Levenshtein, 1966) of reference and hypothesis transcriptions. The WER is thus an error-enumeration based metric which, for its final score, considers every error as equally important. The error importance measure associated to WER is then naturally the occurence count of each error.

When ASR is a first step in a more complex task, such as NER, automatic translation or language understanding, numerous studies shown that the WER is not always well correlated to the performance of the overall task, for example, (Garofolo et al., 2000) in the context of an informa-

tion retrieval task, (He et al., 2011) in the context of speech translation and (Wang et al., 2003) in the context of spoken language understanding.

Some alternative metrics to the WER have been proposed. (Miller, 1955) proposed to measure the loss of information caused by ASR errors and called this metric Relative Information Loss. It is a stochastic based measure which uses the difference of entropy between the hypothesis words as such and in the context of the reference.

Word Information Lost (WIL) has been introduced in (Morris et al., 2004) as an approximation of RIL. For high error rates (Morris et al., 2004) and (McCowan et al., 2004) found that RIL and WIL can be appropriate. Other ASR evaluation measures, inspired by RIL, were proposed. In (McCowan et al., 2004) the authors proposed to adapt the standard metrics used for information extraction (precision, recall and f-measure) to measure the loss of information caused by the ASR errors. The general idea consists in computing the recall and precision at the word level following the alignment between hypothesis and reference produced when computing the WER. There, the ASR is seen as an information extraction problem when the word is the information to find.

In (Garofolo et al., 1999) the authors described 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 Information Retrieval (IR) results was higher than for the WER. One possible cause is that NE-WER ignores inserted or substituted words outside of NE which cause false alarms in the downstream IR. In (Ben Jannet et al., 2015b), a new metric, specifically developed for the context of evaluating ASR systems for a named entity recognition task, was proposed. That metric, ATENE, is based on a probabilistic model that estimates the risk of ASR errors inducing downstream errors in the named entity detection. The metric achieved a higher correlation than WER and NE-WER between the performance in named entities recognition and in automatic speech transcription. This higher correlation is also reported when comparing it to the WIL, and to the triplet P, R and F-measure metrics in (Ben Jannet, 2015).

Having a metric that allows to estimate the quality of an ASR system given a specific task is interesting but doesn't necessarily allow to obtain a list of the most important and frequent errors. However such a list is very important to understand the problems and even improve the ASR system (Dufour and Esteve, 2008). So we not only want a count of elementary errors but also a classification of these errors according to teir possible impact and their ranking given their relative importance with respect to the task, namely named entity recognition. The general metrics RIL and WIL do not provide a quantification of the impact of specific errors. They can only give a general overview of the quality of an ASR system. The seriousness of errors can be estimated when WER, NE-WER and ATENE are used.

In this work, we are interested in establishing a ranking lists of ASR errors according to WER, NE-WER and ATENE and in studing this lists to identify which metric give a more relevent ranking given the application case of NER from ASR output.

3. Proposition

We propose to establish lists of ASR errors and to rank the individual errors according to different evaluation metrics, the widely used WER and two metrics NER-context specific ones which are NE-WER and ATENE. We will first present a general overview of the three metrics which are the basis of our work, then we will describe how we built and ranked the error lists.

3.1. Evaluation metrics

3.1.1. Word Error Rate :

The main ASR metric used to evaluate ASR output in open domain is the WER. It consist to compare a manual transcription (the reference) to ASR transcription (the hypothesis). This comparison is done by applying a Levenstein alignment (Levenshtein, 1966) which project hypothesis on reference allowing to detect ASR errors which are : (D)eleted, (I)nserted and (S)ubstituted words. The WER consist then to estimate the rate of error regarding the number of word (N) to be reccognised in the reference.

$$WER = \frac{S + D + I}{N} \tag{1}$$

3.1.2. Named Entity Word Error Rate

The NE-WER was introduced in order to create a metric more adapted to case of named entities extraction from ASR output.

It is built similarly to the WER, on a Levenstein alignment of reference and hypothesis, but it counts errors only on the named entity spans. NE-WER is given by equation 2, where D_{NE}, I_{NE} and S_{NE} are the numbers of deleted, inserted and substitued words belonging to named entities, and N_{NE} is the total number of words belonging to named entities in the reference.

$$\text{NE-WER} = \frac{D_{NE} + I_{NE} + S_{NE}}{N_{NE}} \tag{2}$$

3.2. ATENE measure

The ATENE measure is fully described in (Ben Jannet et al., 2015b). It aims at quantifying the impact of the errors on the named entity detection by measuring how harder it became to identify entities given the differences between hypothesis and reference by comparing an estimated likelihood of presence of entities. It is based on a maximum entropy classifier to estimate the likelihood. Two sub-metrics are built, one that tries to measure the impact of the ASR errors on the risk of missing, missclassifying or establishing wrong boundaries for the entities, and the other estimating the risk of false alarms.

The word are labelled on whether they are in an entity and of which type it is. To measure the difficulty of distinguishing the correct answer, the *margin* is computed. That margin corresponds to the difference in probability between the reference label $P(\hat{Y})$ and the probability of the most likely incorrect label $\max_{Y \neq \hat{Y}} P(Y)$.

$$M(X) = P(\hat{Y}|X) - \max_{Y \neq \hat{Y}}(P(Y|X)) \tag{3}$$

where X is the vector of features at a given position in the text. The features are the words, prefixes, and suffixes in a [-2;+2] words window. They have been chosen for their simplicity and their use by all NER systems. The estimation of the change in difficulty is established by computing 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.

$$\Delta M(X_A, X_R) = M(X_A) - M(X_R) \quad (4)$$

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

Concerning the ASR errors and their impact on NER, two aspects have to be considered: (1) in an entity zone, ASR errors can cause NER systems to miss or misclassify the entity and (2) outside of an entity zone, ASR errors can cause NER systems to detect non existing entities. This leads to two elementary measures:

- $ATENE_{DS}$ for the case where the most likely label is an entity where the reference label is not or the most likely label is of a different type. It computes the difference between the margin (ΔM) at the start and the end of every entity and the arithmetic mean of all these values is computed to get a global $ATENE_{DS}$ score.
- $ATENE_I$ for the case where a label of an entity presence is the most likely in a non entity zone. It this case, the most negative ΔM in the non entity zone estimates the risk of an insertion; and if there is no negative ΔM then the score is set to 1 (no risk to have an error).

3.3. ASR errors ranking

Our aim being the creation of an error list ranked by their seriousness for the NER task, the first step is then to generate an error list. The Levenstein alignment (Levenshtein, 1966) used to calculate WER and NE-WER allow us to identify ASR errors.

That alignment provides a list of errors of three basic types, insertion, deletion and substitution. Counting these errors allow to estimate their importance from the point of view of the WER since each count, divided by the number of words in the reference, is a direct increment to the final WER score.

In this work, we decided to keep the error-list as-is but to compute the weight of each error through its impact on the ATENE metric. Each entry in the error list is taken one by one. It can be either a deletion, a substitution or an insertion. The idea is then, for each instance of that error, to compute the impact of the instance, then sum them together. For each instance, the first step is to apply the transformation that the error defines to the reference text, to get the hypothesis text. This transformation also provides two informations: an alignment between reference and hypothesis, and an influence zone, the words for which the context has changed, according to ATENE's maximum entropy models features. This transformation is illustrated by Figure 1 for the insertions, Figure 2 for the substitutions and Figure 3 for the deletions,

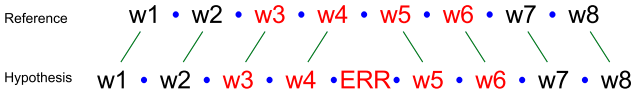

Figure 1: Transformation, alignment and zone of influence for an insertion

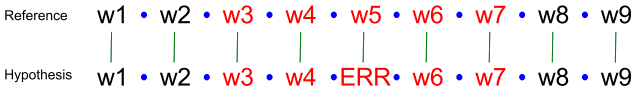

Figure 2: Transformation, alignment and zone of influence for an substitution

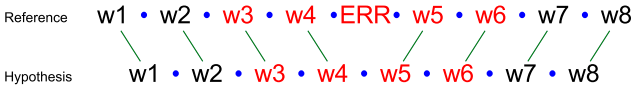

Figure 3: Transformation, alignment and zone of influence for a deletion

The next step is to project the entity boundaries from the reference on the hypothesis. Given the alignment, it can be done implicitely: the boundaries stay anchored on the same words when possible, and when a boundary word is lost due to a deletion the entity size is reduced in the hypothesis. In addition, if an insertion happens at the boundary of an entity, the entity is not extended to include the inserted word.

The final step is then to compute the variation in ATENE, knowing that the probabilities the models compute can only change on the words in the zone of influence. ATENE computes three kind of values:

- The margin difference between hypothesis and reference on the word at the start of an entity
- The margin difference between hypothesis and reference on the word at the end of an entity
- The difference between the minimum margins in the complete out-of-entity spans in the hypothesis and in the reference

For the first two cases, the computation is easy: the entities with one or more boundary in the zone of influence have been projected in the previous step. The margins can be computed on these words with the old and the new contexts and the difference gives the difference in $ATENE_{DS}$. Any boundary falling outside of the zone of influence will have identical context for the original and projected boundary word, and the difference will then be zero. Some words are both a start and an end of entity. In that case both contributions must be computed and added. Also, when an entity is lost, the margin on the hypothesis is set to zero.

The last case is a little more complex, since the entire span must be taken into account. The out-of-entity words in the zone of influence must be followed until reaching the next entity or the limit of the sentence. The minimum of the margin on all these words must be computed, for both the reference and the hypothesis. It needs to be computed only once for the worsd outside of the zone of influence, since

	Test
Words	115,803
Entities	5,933

ASR-1	ASR-2	ASR-3	ASR-4	ASR-5	ROVER
22,3	25,7	26,6	30,4	36,7	28,68

Table 1: Description of the ETAPE corpus and performance of ASR systems obtained during the ETAPE evaluation campaign given in terms of WER.

their context does not change between reference and hypothesis. The margins are then clamped to 1 if they are positive, left as-is if negative, and the difference computed. That gives the $ATENE_I$ difference.
Finally, the two contributions must be combined. The computation of $ATENE_{DS}$ requires computing the mean betwen the margin difference of the beginning and the ending word on each entity. As such, the final summation of the contributions must be divided by two. The final computation of $ATENE$ requires dividing $ATENE_{DS}$ by the number of entities and $ATENE_I$ by the number of inter-entity segments. These numbers being similar, we decided to just add the two values together. The final sum for each error gives its absolute importance for the metric, allowing to sort the list of errors accordingly.
These computations then allow to sort the error lists by their importance. The usual WER evaluation tool, *sclite*, already provides the WER-related list through the *dtl* output (WER list). We created the equivalent list using the ATENE impact as an importance quantifier (Atene list).
In addition, for the purpose of comparisons, we also built two other lists based on occurrence counts. One, inspired by NE-WER, only keeps the error instances that happen inside named entities (In list). The second adds to the In list errors those that are touching an entity, to get the nearest context (Near list).

4. Experiments and Results

4.1. Data

Our experiments are conducted on the ETAPE data (Galibert et al., 2014). This corpus contains 15 radio broadcast, manually transcribed and automatically transcribed by 5 different ASR systems and a ROVER system as summarized in Table 1.

4.2. Methodology

Two different error lists was generated for each metrics (ATENE, WER, In and Near). The first one contain the ten most serious errors and indicates the rank obtained for each metric (List-10). The second one is the built using the same process with the 100 most serious errors (List-100). First, we can observe that there is an overlap between the error lists generated by the different metrics. Indeed, if all the metrics gave a list of different errors then there should be $10\,(errors) \times 6\,(systems) \times 4\,(metrics) = 240$ entries in the fusion of all the List-10. But only 47 are present. The same observation can be made on the fusion of all the List-100. There could be at most $100 \times 6 \times 4 = 2400$ entries but there are only 733. We want to compare those lists following these two hypothesis:

- The lists generated for one system by the four metrics are different. If this hypothesis is verified, the impact of the errors measured on the same system is not the same according to the metric used;
- The ranking of the errors given its seriousness is equivalent regardless of the system. If this hypothesis is verified, the metric is consistent.

To compare the ranked lists and then verify those hypothesis, we use a rank correlation coefficient, specifically the Spearman's ρ (Spearman, 1904). Spearman's correlation is reflecting the degree of concordance and discordance on the rank scale. This measure gives values between -1 and +1 indicating the power of correlation between the two tested variables. If the value is high ($\geq 0,8$), this means that the order of the errors is the same; If the absolute value is low, this means that this is not the same errors that are evaluated as serious. A negative value indicates that the lists are in reverse order.
An other way to evaluate the quality of those lists is to analyze the lists themselves (their content) given the objective (here impact on NER). This analysis, even if partial, should give interesting insights.

4.3. Lists comparisons

The rank correlations are calculated by pairs for each system and are presented in matrices in Figure 4 for the List-10 and Figure 5 for List-100. The highest a correlation is ($R > 0,8$), the greener the box is. The lowest a correlation is ($R < 0,3$), the redder the box is.
First, we want to observe whether the metrics are consistent given the systems. The objective is to verify if the lists generated by one metric are equivalent for every system.
As shown in Figure 4 and Figure 5, the mean correlation obtained with the WER is 0.96 ($\sigma = 0.04$) for List-10 and 0.91 ($\sigma = 0.06$) for List-100. For ATENE, the mean correlation is 0.85 ($\sigma = 0,10$) for List-10 and 0.83 ($\sigma = 0.09$) for List-100. For Near, the mean correlation for List-10 is 0.84 (($\sigma = 0.14$), but it decreases to 0.72 ($\sigma = 0,16$) for List-100. For In, the mean correlation is 0.66 ($\sigma = 0.22$) for List-10 and 0.64 ($\sigma = 0.23$) for List-100.
We can conclude that the lists obtained with WER and ATENE are the most consistent with respect to the different systems, having a mean correlation above 0.8. However the metric In is the one that seems to be less consistent with a mean correlation lower than 0.7. The WER and ATENE metrics estimate the impact of errors in a similar way from one system to another. Changing the ASR system should not change the error list. These two metrics are not only consistent but also robust.
Our second question is to verify whether the error lists are different given the metrics in order to assess their complementarity. Concerning WER, the mean correlation with ATENE is 0.10 ($\sigma = 0.19$) for List-10, 0.03 ($\sigma = 0.23$) with In and 0.64 ($\sigma = 0.10$) with Near. The same trend is observed with List-100: its mean correlation is -0.17

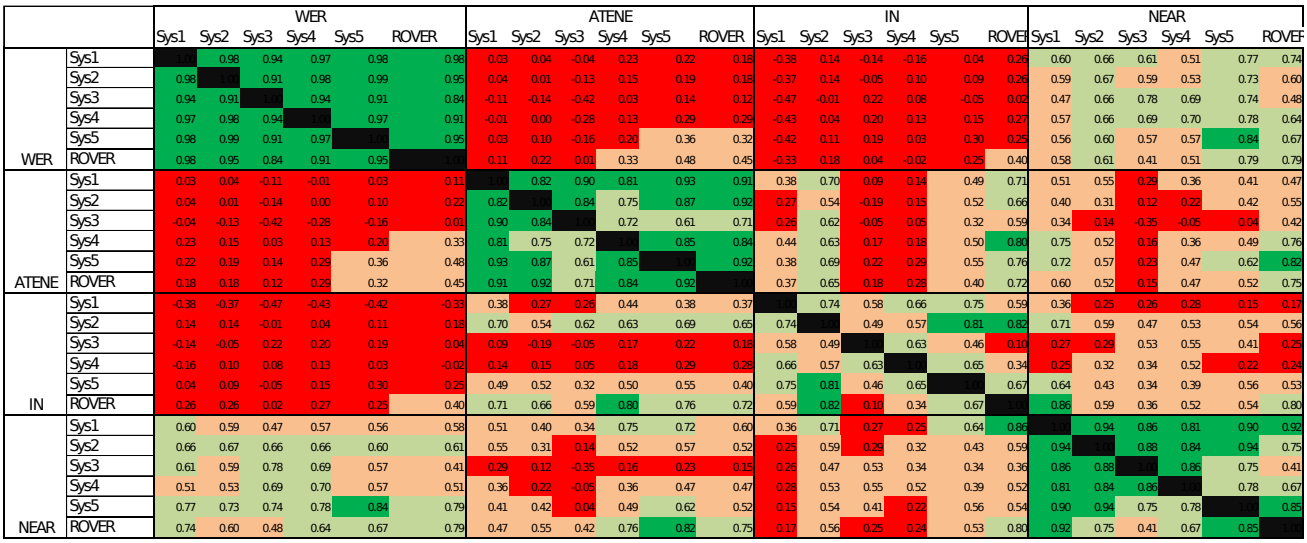

Figure 4: List-10: Correlation matrix of the different measures, WER, ATENE, IN and Near, for all the 5 ASR systems and the rover.

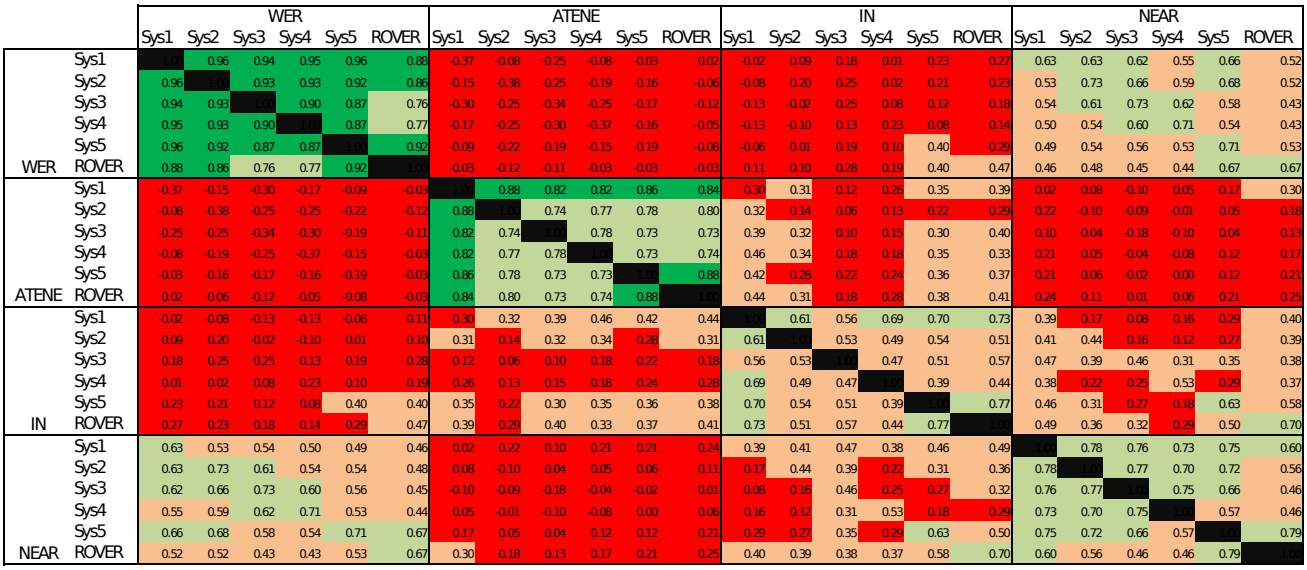

Figure 5: List-100: Correlation matrix of the different measures WER, ATENE, In and Near for all the 5 systems and the rover.

($\sigma = 0.09$) with Near. The very low correlation between WER and In shows that the words within the entities are very specific in relation to the general language. However, WER correlates better with a metric that highlights errors around EN such as Near. This is likely due to the fact that adding the words around the EN dilutes their specificity and makes them more similar to the global WER list.

Observing the correlation between ATENE and other metrics, a very low correlation with WER (0.10) is observed. This shows that these two metrics highlight very different errors. Its mean correlation with Near is a little bit higher (0.40, $\sigma = 0.11$) and In (0.41, $\sigma = 0.25$) for List-10. ATENE seems to provide very different information than the WER due to its consideration of the NER task. It is closer to what happens to the entity words, but goes further by taking the context into account.

4.4. Qualitative analysis

Table 2 gives some examples extracted from the lists. It includes the first ranked error for every type (Deletion, Insertion and Substitution) and for every list (Atene, WER, In, Near).

As we can see, the first Atene deletion is the preposition *à*. This preposition is in general a good marker for named entities. In the ETAPE reference, 68% of its occurences are before a named entity. Losing such a word ends up being quite important for the detection. While WER sees a large number of occurences, unsurprisingly for a single-phone word, the In and Near-derived lists fail at noticing its specific importance for named entities (respectively ranked at 16th and 57th position).

The first insertion, *dix*, a number, is even more significant. It is always inside an entity (amount or time/date) but the other metrics fail at noticing the importance of its insertion,

Deletions				
Atene	Word	WER	In	Near
1	à	7	16	57
873	il	**1**	9	7
9	de	5	**1**	**1**

Insertions				
Atene	Word	WER	In	Near
10	dix	673	-	2220
8112	et	**14**	16	**4**
8184	des	87	**12**	23

Substitutions					
Atene	Ref	Hyp	WER	In	Near
2	de	deux	124	98	48
4863	il	qui	**38**	-	-
8	deux	de	116	**4**	**18**

Table 2: Extract from the lists, with the rank with the proposed method (Atene), and the ranks for WER, for errors inside entities (In), and for errors inside or touching entities (Near). In bold is the first-ranked case for each sorting method.

which is certain to create a false positive in the NER system (position 623 for WER and 2220 for Near; it does not appear for In because the insertion always happens outside of a reference entity zone).

The first substitution, *de*, the determiner, into *deux*, the number, suffers from a similar problem with traditional metrics: it has a near certain chance of producing a false positive or of breaking a large entity in multiple parts, but is not considered that important for the other metrics, comparatively (position 124 for WER, 98 for In and 48 for Near). The inverse substitution on the other hand is noticed by the In and Near sorting methods (4th and 18th ranks), since they happen inside entities. It corroborates the Atene study (Ben Jannet et al., 2015a) which shows that NE-WER is nearly as good as Atene for taking into account entity deletion or incorrect typing. The important point though is that the Atene list also gives that error a good rank (8th).

The most important deletion for WER, the personal pronoun *il*, and the most important insertions *et*, the coordinating conjunction and *des* the determiner, are indeed not very impacting for NER detection, no matter where they happen. But as small monophonic words they happen often, hence their high ranking for everything but Atene.

Finally, the *de* deletion is important in every list, which is as expected since it happens often and its loss can easily make the NER system break an entity into two.

As a result, that small analysis shows that the Atene-motivated error list has a higher potential as a tool to make systems better. And finally it is interesting to note that the important errors can be hard to fix in the ASR system but, once identified, can be compensated for in the NER system, especially if it's a symbolic one.

5. Conclusion

In this paper we presented a methodology to generate list of ASR errors ranked given their impact on the task of Named Entity Recognition. This methodology is based on an alignment between the words of the reference and the hypothesis as provided by the computation of the WER metric. In this work, we keep the WER error-list as-is and also compute the weight of each error through its impact on the ATENE metric. Moreover, we also built two other lists based on occurrence counts. The first one keeps the error instances that happen inside named entities (In list) and the other one keeps those and the ones that are touching an entity, to get the nearest context (Near list).

That methodology was applied to the data provided by the ETAPE evaluation campaign. We provided a comparison of the different lists obtained through these different methodologies. The correlation measures we have carried out show that the WER and ATENE metrics are consistent for each system and thus provide information that is largely independent of one ASR system. They are also fairly correlated, which means that they give different information on the seriousness of ASR errors. A more detailed analysis shows that the estimation of the error impact given by ATENE is more adapted to the NER task than the estimation based only on the frequency. This means that we managed to provide an ordered list of errors having an impact for the downstream system. The next step will be leveraging the lists to make systems better. While correcting the ASR output is probably not doable, the errors being of the "hard" category, we expect that the lists are usable to make the NER systems more robust the kind of errors they should expect to encounter.

Acknowledgements

This work was funded by the ANR VERA project (ANR 12 BS02 006 04)

Adda-Decker, M. (2006). De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux. In *Proc of JEP*, Dinard, France.

Ben Jannet, M. A., Galibert, O., Adda-Decker, M., and Rosset, S. (2015a). How to evaluate asr errors impact on ner? In *Errare Workshop*, Sinaia, Romania, September.

Ben Jannet, M. A., Galibert, O., Adda-Decker, M., and Rosset, S. (2015b). How to evaluate asr output for named entity recognition? In *Interspeech*, Dresden, Germany, September.

Ben Jannet, M. A. (2015). *Évaluation adaptative des systèmes de transcription en contextes applicatifs*. Ph.D. thesis, Université Paris Sud, octobre.

Boháč, M., Nouza, J., and Blavka, K. (2012). Investigation on most frequent errors in large-scale speech recognition applications. In *Text, Speech and Dialogue*, pages 520–527. Springer.

Comas, P. R. and Turmo, J. (2009). Robust question answering for speech transcripts: Upc experience in qast 2009. In *Working Notes of CLEF 2009*, Corfou, Grèce, October.

Dinarelli, M. and Rosset, S. (2011). Models Cascade for Tree-Structured Named Entity Detection. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 1269–1278, Chiang Mai, Thailand, November.

Dufour, R. and Esteve, Y. (2008). Correcting asr outputs: Specific solutions to specific errors in french. In *Spoken Language Technology Workshop, 2008. SLT 2008. IEEE*, pages 213–216, Dec.

Galibert, O., Leixa, J., Adda, G., Choukri, K., and Gravier, G. (2014). The ETAPE speech processing evaluation. In *LREC*, Reykjavik, Iceland.

Garofolo, J. S., Voorhees, E. M., Auzanne, C. G., Stanford, V. M., and Lund, B. A. (1999). 1998 trec-7 spoken document retrieval track overview and results. In *Broadcast News Workshop'99 Proceedings*, page 215. Morgan Kaufmann Pub.

Garofolo, J. S., Auzanne, C. G., and Voorhees, E. M. (2000). The trec spoken document retrieval track: A success story. *NIST SPECIAL PUBLICATION SP*, 500(246):107–130.

Ghannay, S., Estève, Y., and Camelin, N. (2015). Word embeddings combination and neural networks for robustness in asr error detection. In *EUSIPCO*, Nice, France.

Goryainova, M., Grouin, C., Rosset, S., and Vasilescu, I. (2014). Morpho-syntactic study of errors from speech recognition system. In *LREC*, Reykjavik, Iceland, May.

He, X., Deng, L., and Acero, A. (2011). Why word error rate is not a good metric for speech recognizer training for the speech translation task? In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, pages 5632–5635. IEEE.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710.

Lippmann, R. P. (1997). Speech recognition by machines and humans. *Speech Communication*, 22(1):1–15.

McCowan, I. A., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., and Bourlard, H. (2004). On the use of information retrieval measures for speech recognition evaluation. Technical report, IDIAP.

Miller, G. A. (1955). Note on the bias of information estimates. *Information theory in psychology: Problems and methods*, 2:95–100.

Morris, A. C., Maier, V., and Green, P. (2004). From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *INTERSPEECH*.

Rena Nemoto, I. V. and Adda-Decker, M. (2008). Speech errors on frequently observed homophones in french: Perceptual evaluation vs automatic classification. In *LREC*, Marrakech, Morocco, May.

Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication*, 49(5):336–347.

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101.

Vasilescu, I., Adda-Decker, M., and Lamel, L. (2012). Cross-lingual studies of ASR errors: paradigms for perceptual evaluations. In *LREC*, Istanbul, Turkey.

Wang, Y.-Y., Acero, A., and Chelba, C. (2003). 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*, pages 577–582. IEEE.