How to evaluate ASR output for named entity recognition?

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, these 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 metric to estimate the quality of automatic transcriptions within their larger applicative context.
We introduce a new measure to evaluate ASR in the context of named entity recognition, which makes use of a probabilistic model to estimate the risk of ASR errors inducing downstream errors in named entity detection. Our evaluation, on the ETAPE data, shows that ATENE achieves a higher correlation than WER between the performances in named entities recognition and in automatic speech transcription.
Index Terms: speech recognition, ATENE, named entity recognition, metric

1. Introduction

Tremendous progress has been achieved during the last decades, for example in open-vocabulary continuous speech recognition (see [1] or [2]) or in robustness against speaker variation and noisy environment (see [3] or [4]). ASR transcriptions have become accurate enough to be efficiently used in applications such as speech-to-speech translation, spoken language dialog systems or spoken information retrieval [5]. However, ASR still produces errors, mainly due to challenging acoustic conditions, out-of-vocabulary words or language ambiguities. This errors are of varying importance with respect to the application.

ASR systems which are embedded in complex spoken language applications ask for a new metric which modulates transcription errors with respect to their impact on the overall application performance. This leads to a loss of interest in the mere WER metric, which considers each error to carry equal weight. Our objective is then to provide a metric which measures the fitness of the ASR output to the overall task rather than a general ASR-centered metric such as WER.

In this paper, we consider the case of ASR systems in a speech-based Named Entity Recognition (NER) task. In Section 2, we discuss the existing measures to evaluate ASR transcriptions. After a description of the considered NER task in Section 3, a new measure *Automatic Transcription Evaluation for Named Entity* (ATENE) is described in Section 4. Section 5 provides results with a comparison between standard metrics and ATENE with respect to NER system performance along with an analysis of results.

2. Related Work

Speech recognition 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. In order to optimize the performances of their applications developers have to select the most appropriate ASR system considering the global use case.

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 [6] 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 [7]. 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 ([8] in the context of an information retrieval task, [9] in the context of speech translation and [10] 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 [11], 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, in that it considers RIL has two

main problems: (i)it considers systematic words substitutions as correct, and (ii) its implementation is very complex [12]. Word Information Lost (WIL) has been introduced in [13] as an approximation of RIL. For high error rates [13] and [14] 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 [8] 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 [15] 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 [16], the Named Entities Recognition task has become a critical step in numerous language processing applications [17]. 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 exist. For this study we concerned ourselves with the Quaero [18, 19] 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. That schema has been used in two evaluations, in 2010 within Quaero [20] and in 2014 with the open campaign ETAPE [21]. We use the ETER (Entities Tree Error Rate) [22] to evaluate the task in both manual and automatic transcription conditions:

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

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 in reference and hypothesis. And N is the total number of entities in the reference. That metric is similar to the Slot Error Rate [23] 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 we use in our evaluations (see Section 5.4, is that the metric can be decomposed in its individual parts, insertions (I), deletions (D) and substitutions (E).

4. Proposition

4.1. 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. To classify the entities the NER systems rely on multiple levels of contextual features such as the words, or parts of speech. The ASR errors modify these features and thus influence the NER decisions. The impact of an error depends on its nature, but also on the context in which it happens. That makes it important to take into account the whole contextual information in the building of the measure.

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. Likelihood estimation requires a statistical representation of named entity presence on proper transcriptions. We propose to use a simple maximum entropy classifier which uses basic features (words, prefixes and suffixes) common to symbolic and stochastic approaches in NER to obtain our statistical representation. 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.

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 $\max_{Y \neq \hat{Y}} P(Y)$.

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

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.

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

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

4.2. Entity projection

Computing ΔM requires being able to align positions between the reference manual transcription and the ASR output. We do that alignment by extending the approach defined in [24]. 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. The conversion is done through a pronunciation dictionary. The procedure then gives us positional associations between reference and hypothesis for the words at the start and the end of every entity.

4.3. Elementary measures : $ATENE_{DS}$ and $ATENE_I$

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

a wrong label probability going up. Both cases do not need to be distinguished. We measure the risk of deletion or substitution by computing ΔM at the start and the end of every entity. We then take the arithmetic mean of all these values to get a global $ATENE_{DS}$ score.

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

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. 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 provides a decomposition of the ASR hypothesis in segments that are alternatively inside and outside an entity. It is thus possible to collate all word hypothesis and reference 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 in words changes between reference and hypothesis with the ASR insertions and deletions. Statistics on our development data has shown that for inter-entity spans insertion errors counts were 0 or 1 in more than 90% of the cases. So we propose for each segment to estimate the risk of having at least one insertion error by finding the lowest correct detection margin for all words in the span. Noting $\emptyset$ the out-of-entity label:

$$M_O(S) = \min_{X \in S} \left(P(\emptyset|X) - \max_{Y!=\emptyset} P(Y|X) \right) \quad (8)$$

Where S is an outside-entities segment. Since we aim at detecting errors only, we decided to take the margin into account only when it is negative, e.g. when an error seems possible. Otherwise the margin is sent to 1:

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

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

$$\Delta M_O(S) = M'_O(S_A) - M'_O(S_R) \quad (10)$$

Where S_A *and* S_R are matched segments in ASR transcript and reference out side named entities.

$$ATENE_I = \frac{\sum_{i=1}^{N_S} \Delta M_O(S_i)}{N_S} \quad (11)$$

Where N_S is the number of segments out side named entities in hypothesis. We expect a large negative value to be correlated to a high rate of insertion, e.g. a low precision.

4.4. Final score

In order to compute a unique score to the whole ASR transcriptions, 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 in a compatible manner and the simple mean is the correct method.

$$ATENE = -100 \frac{ATENE_{DS} + ATENE_I}{2} \quad (12)$$

The final -100 multiplier is added to reach a behaviour similar to an error rate, e.g. 0 is the best possible value, and the higher the value the worse the score is.

5. Experiments and Results

5.1. Data Description

For our experiments we used the QUAERO [25] and ETAPE [26] data sets. Both corpora are fully annotated in named entities according to the [18] guidelines. Both corpora has been used for an evaluation of NER on automatically transcribed speech. As a result multiple automatic transcriptions, and multiple NER runs from very different systems are available in these data set. Table 1 provides some statistics for the corpus in terms of number of words and named entities. The

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

	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

ETAPE test data set includes six automatic transcriptions provided by five different ASR systems plus a rover, all annotated by seven NER systems. Two of the NER system are based on a symbolic approach and the others on stochastic or hybrid approaches. It is composed of 15 files, each one being an instance of a broadcast news or broadcast conversations program.

5.2. Implementation

The methodology proposed in Section 4 relies on statistical models predicting the presence of NE 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 [27]. Standard features, being a subset of both symbolic and stochastic approaches, were used:

- Uni- 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
- Uni- and bi-grams of Part of speech (POS) in a [-1,+1] window of around the target word

To study the impact of the features and potential bias, four different models are trained. The first one (baseline) is based on the uni- and bi-grams of words, the second one adds the prefixes and suffixes, the third one adds to the baseline model the POS and finally the fourth one is based on all these features. Capitalization is also a feature used by NER systems but not all ASR systems provide one. In our test data, only one ASR system offers a capitalized transcription. So we trained a variant of each previous models taking that into account.

Estimating $ATENE_{DS}$ requires two models, one predicting the start of entities with their type and one predicting the end. So the first model predicts either B-*type* or O as in outside. The second model predicts either E-*type* or O. Estimating $ATENE_I$ on the other hand only detects the presence of an

Table 2: Mean Spearman correlation between ETER and scores given by WER, NE-WER and $ATENE$ on ETAPE challenge data.

		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	*Baseline*	0.70	0.79	0.47	0.64	0.73	0.80	0.71	0.69
	Base+pref+suf	0.74	0.82	0.45	0.66	0.77	0.78	0.73	0.71
	Base+pos	0.71	0.78	0.42	0.66	0.74	0.78	0.73	0.69
	Base+pref+suf+pos	0.74	0.80	0.43	0.67	0.78	0.80	0.74	0.71
	Baseline+capital	0.39	0.64	0.59	0.53	0.68	0.69	0.66	0.60
	Base+pref+suf+capital	0.43	0.75	0.66	0.56	0.71	0.77	0.74	0.66
	Base+pos+capital	0.37	0.69	0.57	0.49	0.65	0.69	0.67	0.59
	Base+pref+suf+pos+capital	0.47	0.74	0.56	0.61	0.69	0.77	0.73	0.65

entity. We decided to make it predict I-*type*, as in Inside, or O. That makes that last model consistent with the others.

5.3. Evaluation methodology

Comparing ASR systems based on the quality of their transcript 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 [28]. 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 [29]. We propose to use Spearman ρ because it handles the case where ties ranks are present. We noticed though that using Kendall's τ did not change the relative results, only the absolute values.

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

5.4. Results

As shown in Table 2, the ATENE measure, based on the models without the capitalisation features, (see first 4 rows), correlates better than the WER and the NE-WER with the results of most NER systems. The only exception is with the NER-1 system, and not by much. The mean correlation (see right column) is also better with the ATENE measure than with the two other ones. If the baseline model provides good results, those are better when using morphological features in addition to the uni- and bi-grams of words. It is interesting to note that the NER-5 system, having a mean correlation of 0.74 is a pure symbolic one. This shows that the measure's stochastic roots do not bias against symbolic systems.

We observe a very small change in correlation results when using the first four models which include a different features combination. This observation reinforces our proposal and shows that our measure can be used even without knowing the features used by the target NER system(s).

The correlation decreases a little when using the models including capitalization as features. This is because only one ASR system provides upper-cases. The output of this system is promoted when capitalization is taken into account by the classifier, biasing the results in favor of this system.

The ATENE measure correlates better with the NER-3 results when capitalization is taken into account, while the opposite observation is done for NER-1. That suggests that upper-cases are an important features for NER-3 and not for NER-1. For the other NER systems the difference is less important. This can be due to the use of different version of their systems to handle both conditions, with and without upper-cases.

6. Conclusion

This paper addressed the issue of evaluating the quality of ASR transcriptions in a complex NLP task combining ASR with NER. Standard metrics, such as WER and NE-WER show a relatively low correlation between ASR and NER performances using Spearman's ρ 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 measure, a major weakness consists in not 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 ASR mistakes. Our measure takes into account the different kind of errors that ASR transcript may entail in downstream NER systems. The merits of ATENE were tested by comparing it to WER and NE-WER on ETAPE benchmark data as produced by several different ASR and NER systems. The ATENE measure achieves a higher correlation with NER results than do WER and NE-WER, thus showing the added value of this new measure. Moreover we provide a deep analysis of the impact of the features used for building the model. This analysis shows that the measure's stochastic roots do not bias against symbolic systems.

Future work includes the optimization of the ASR system settings with respect to our ATENE measure.

7. Acknowledgements

This work was funded by the ANR VERA project (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).

8. 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 subword 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 *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4681–4684.

[5] L. Lamel, "Multilingual speech processing activities in quaero: Application to multimedia search in unstructured data," in *The Fifth International Conference: Human Language Technologies - The Baltic Perspective*, 2012, pp. 1–8.

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

[7] 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.

[8] 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.

[9] 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 *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5632–5635.

[10] 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.

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

[12] 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.

[13] 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.

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

[15] 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.

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

[17] 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

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

[19] 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

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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

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

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