INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Investigating the Effect of ASR tuning on Named Entity Recognition

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

[1]LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
[2]LNE, F-78190 Trappes, France
[3]LPP–CNRS UMR 7018, Université Sorbonne Nouvelle
{first.last}@limsi.fr, {first.last}@lne.fr
email@address

Abstract

Information retrieval from speech is a key technology for many applications, as it allows access to large amounts of audio data. This technology requires two major components: an automatic speech recognizer (ASR) and a text-based information retrieval module such as a key word extractor or a named entity recognizer (NER). When combining the two components, the resulting final application needs to be globally optimized. However, ASR and information retrieval are usually developed and optimized separately. The ASR tends to be optimized to reduce the word error rate (WER), a metric which does not take into account the contextual and syntactic roles of the words, which are valuable information for information retrieval systems. In this paper we investigate different ways to tune the ASR for a speech-based NER system. In an end-to-end configuration we also tested several ASR metrics, including WER, NE-WER and ATENE, as well as the use of an oracle during the development step. Our results show that using a NER oracle to tune the system reduces the named entity recognition error rate by more than 1% absolute, and using the ATENE metric allows us to reduce it by more than 0.75%. We also show that these optimization approaches favor a higher ASR language model weight which entails an overall gain in NER performance, despite a local increase of the WER.

Index Terms: automatic speech recognition, tuning, named entity recognition

1. Introduction

Named entity recognition (NER) from spoken documents requires the combination of NER and ASR (Automatic Speech Recognition) systems. Hence, the overall system performance depends on both components. In this paper we investigate ASR–NER coupling issues from an ASR perspective.

Usually ASR systems are tuned empirically, in isolation, by selecting the parameters that minimize the WER (Word Error Rate). In order to establish if certain ASR tuning practices had a negative effect on NER performance, we firstly investigate the relationship between ASR performances in terms of WER and the effect of varying the ASR tuning methodology on named entity recognition error rate. We also investigate the use of other ASR evaluation metrics such as the NE-WER (Named Entity Word Error Rate) and the ATENE (Automatic Transcriptions Evaluation for Named Entity) for ASR tuning and compare the results with those obtained using the standard metric (WER). During this tuning step, we also use the NER system as an "oracle" to estimate what the best possible result could be.

The remainder of the paper is organized as follows: after a presentation of related work (Section 2), the experimental setup including task, data and models used in our study is described in Section 3. Next, Section 4 presents the achieved results along with an analysis and discussion. Finally, Section 5 concludes the paper.

2. Related Work

Over the last decades, ASR systems have commonly be tuned to minimize the WER, which represents the ratio of wrongly transcribed words among the total number of words to be recognized. Although, this is perfectly appropriate in a stand-alone approach, one may wonder whether WER minimization is still the best way to tune ASR systems as soon as they are to be coupled with other natural language processing components? Previous work on coupling or embedding ASR within other Natural Language Processing (NLP) modules studied the effects of WER on the global application performance. Thereby [1] in the context of speech translation, [2] in the context of keyword spotting, [3] in the context of named entity recognition, and [4] in the context of spoken language understanding argued that the WER is not always best correlated with the application performance.

In the context of speech translation, the work of [1] and [5] on ASR tuning for speech translation demonstrated that ASR should be optimized considering a translation evaluation metric (e.g. BLEU) instead of the traditional WER. They observed that a larger language model (LM) weight gave better speech translation performance although it locally lead to an increased WER for ASR proper.

In the case of information retrieval (IR) from speech or named entity recognition on speech, we found no work describing a specific ASR tuning study. However we found studies proposing alternative metrics to evaluate ASR in the case of NER from speech. Thereby [6] proposes NE-WER (Named Entity Word Error Rate) and lately [3] proposes ATENE (Automatic Transcription Evaluation for Named Entity).

NE-WER is an error rate calculated only for words belonging to a named entity. In their study [6] authors show that NE-WER correlates better than WER with keyword spotting performance in the case of the TREC evaluation results.

$$\text{NE-WER} = \frac{E_{NE}}{N_{NE}} \quad (1)$$

With E_{NE} is the number of words incorrectly transcribed inside named entities and N_{NE} is the total number of words occurring inside named entities in the reference transcription.

ATENE provides an estimation of the additional difficulty introduced by ASR errors for NER. This estimation is calculated by computing the difference between the probability of having a named entity in the reference and in the ASR hypothesis at the same position. Probabilities are given by a Maxi-

http://dx.doi.org/10.21437/Interspeech.2017-1482

mum Entropy [7] statistical model. In their study the authors show that ATENE correlates better than WER with the NER performance in the case of the ETAPE [8] and QUAERO [9] evaluation campaign data. Equation 2 gives the formula for this metric.

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

$ATENE_{DS}$, given by 3, mesures the risk of entities deletion and substitution introduced by ASR transcriptions. $ATENE_I$, given by 4, mesures the risk of entities insertion (false alarm) introduced.

$$ATENE_{DS} = \frac{\sum_{i=1}^{N} \Delta_p(\text{begin}_i) + \Delta_p(\text{end}_i)}{2N} \quad (3)$$

Δ_p is the difference of probability calculated between words in the reference and in the ASR hypothesis at the beginning and end of entities. Thanks to the MaxEnt model, the probabilities take into account the contextual information around the target words. N is the number of named entities in the reference.

$$ATENE_I = \frac{\sum_{i=1}^{N_S} \Delta_{PS}(S_i)}{N_S} \quad (4)$$

Δ_{PS} is the difference between the probability of having at least one false alarm in a text segment between two named entities. N_S is the number of such text segments.

Previous work on ASR based natural language processing applications suggests that it may be beneficial to specifically tune the ASR system to optimize the overall embedding application, and that the WER is not necessarily the best metric in this case.

3. Task and data

3.1. Named entity task description

Since its creation in the MUC conferences [10], the Named Entities Recognition task has become a crucial step in numerous language processing applications [11]. The task consists in detecting, classifying and decomposing all mentions of named entities which are, in a loose approximation, objects and concepts of the real world the discourse is referring to. Numerous annotation schemes with varying degrees of complexity and coverage exist. This study makes use of a hierarchical annotation scheme [12, 13], which was originally developed for French evaluation campaigns, but also extended to English and other languages. It proposes a structural organization of complex named entities and provides a better coverage than most other schemes. ASR tuning methods validated using this complex annotation scheme should easily port to simpler Named Entity (NE) schemes.

The taxonomy involves seven classical named entity classes: person, location, organization, function, product, temporal expression and amount. Annotation rules include two levels, the first level consists of a full entity classification and the second one of the entities' decomposition into differently typed slots. The taxonomy is hierarchical with a number of sub-types, and the annotation recursive, entities may be included in other entities. Figure 1 illustrates this extended annotation.

This annotation scheme has been used during two evaluation campaigns, in 2010 within the French Quaero [14] campaign and in 2014 within the ANR French open ETAPE [15] campaign. ETER (Entities Tree Error Rate) [16] is the corresponding metric to evaluate the task in both manual and automatic transcription conditions:

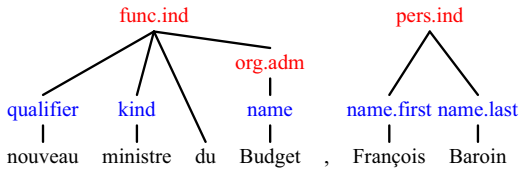

Figure 1: *Multi-level annotation with types (func.ind, pers.ind and org.adm) and components (qualifier, kind, name, first.name and last.name).* new minister of budget, François Baroin.

$$ETER = \frac{I + D + \sum_{(e_r,e_h)} E(e_r, e_h)}{N} \quad (5)$$

where I and D represent the number of inserted and deleted entities determined through an alignment of reference and hypotheses annotations. $E(e_r, e_h)$ indicates the sum of classification and decomposition errors in aligned reference and hypothesis entities. And N is the total number of entities in the reference. The ETER metric is similar to the Slot Error Rate [17] while providing a more elaborate classification/decomposition error estimation. It can be seen as an error enumeration metric, conceptually similar to WER. An interesting property (and that can be seen in the formula) is that the metric can be decomposed in its individual parts, insertions (I), deletions (D) and substitutions (E).

Theoretically, optimizing a full ASR–NER application requires a joint minimization of ASR–NER metrics. In practice, this is difficult to implement and even often impossible, as the composing ASR and NER subsystems may either come from different organizations or (and) be developed at different timelines. In the next section, we propose an alternative to avoid or at least to alleviate this problem. If ASR tuning can be carried out using a metric which better correlates with the final NER metric than standard ASR WER does, the presence of the NER system becomes less crucial.

3.2. Data description

The training data consists of a hundred hours of speech from various broadcast sources which were first provided during the ESTER evaluation campaign [18], and then annotated in named entities during the Quaero[1] project [13]. The development data set consists of approximately 8h20 of speech from various broadcast sources. All data are manually transcribed, and annotated in named entities according to the [12] guidelines.

Table 1 provides some statistics about the QUAERO corpus in terms of number of words and named entities. In this paper we use the ASR dev data, which is roughly half of the NE test data.

3.3. ASR system

We used in this experiments the LIMSI ASR system which is widely described in [19, 20]. It is a state-of-the-art system, and the last stage was changed to generate large lattices. That allows to efficiently generate multiple 1-best outputs with different tuning parameter values.

[1] freely available for research purposes at ELDA, under the reference ELRA-S0349, ISLRN : 074-668-446-920-0.

Table 1: *Overview of the the QUAERO corpus.*

	QUAERO	
	Train	Dev
Words	1 251 586	45 305
Ents.	113 885	3 267

3.4. NER system

The NER system used in our experiment is based on CRF (Conditional Random Fiels). The model was trained on the QUAERO training data set using the WAPITI software [21]. The model is based on the following set of features:

- Words and bi-grams of words situated in a [-4,+4] window of around the target word
- Prefixes and suffixes of words situated in a [-4,+4] window of around the target word
- The presence of capital letter, number or punctuation on the target word

In order to deal with the hierarchical structure of the QUAERO named entity definition, we use a set of combined labels. Each token is associated with a combination of the type and the component labels associated to it. The classical BIO schema was used. This leads to a set of 1700 possible labels.

To check its performance, we applied this model on the ETAPE[2] test data [15] which follows the same NE schema. Our model obtains an Entity Tree Error Rate (ETER) [22] of 41.7% which would rank this system at the 6th position on 11 systems given the results of the evaluation campaign.

4. Experiments, results and analysis

4.1. Objectives

Our aim is to investigate the impact of the ASR tuning and the ASR error metrics on the NER performance. We structured our experiments in two steps:

- studying the impact of the tuning parameter changes on the different metrics we are investigating (WER, NE-WER, ATENE, and final post-NER ETER score) and check which has the best correlation with the final ETER;
- evaluating the final quality of the output when tuning on the different metrics (WER, NE-WER, ATENE), plus, as an informative data point, comparing with the result of a full end-to-end "oracle" tuning with the NER system in the loop.

4.2. Tuning and experimental runs

Two tuning parameters were optimized in our study:

- The language model (LM) weight which balances the scores contributions from the acoustic and language models. The LM weight (LMW) was varied between 0 to 80 in increments of 2.
- The word insertion penalty which was adjusted for each LM weight in order to optimize each metric (WER, ATENE, NE-WER and ETER).

[2]freely available for research purposes at ELDA, under the reference ELRA-E0046, ISLRN : 425-777-374-455-4.

Each time these two parameters are changed on the development set the WER, ATENE and NE-WER are computed on the ASR output. The NER system is then applied to the generated ASR output and the named entities error rate is measured using ETER. This procedure entails a large number of ASR and NER runs.

Since we have both transcriptions and named entity reference annotations only for development corpus of QUAERO, a cross validation was performed by splitting the data into seven parts (around 460 named entities each), using five of them for the tuning and the remaining two for evaluating the result. Five out of seven giving 21 possibilities, all of them were run and the final evaluation scores then correspond to the mean of all 21 individual results.

4.3. Experimental results

4.3.1. Decoder tuning and NER performance

Our first experiment was to study the impact of changing the tuning parameters on our metrics of interest: WER, NE-WER and ATENE. For comparison purposes, we also applied the NER system to the ASR system output to measure the actual final ETER score. We changed the LMW and, for each value, optimized the word insertion penalty to reach the best score for the metric of interest. This gives the results in Figure 2, where each dot represents the mean of all measures obtained for a given LMW on all 21 development sets.

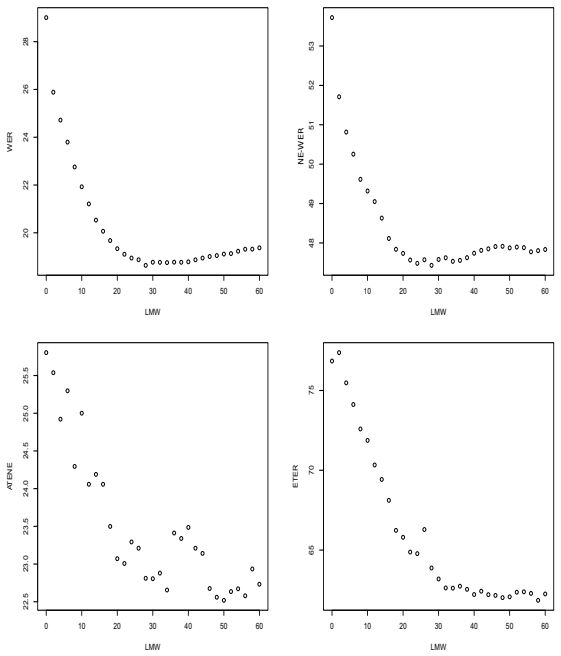

Figure 2: *WER, NE-WER, ATENE and NER performance (ETER) as a function of the LMW during the tuning.*

Taking a look at the two plots on the top of Figure 2 corresponding respectively to WER *vs* LMW and NE-WER *vs* LMW, we can observe that for these two metrics the optimal performances are obtained for a value of LMW around 30, while a

larger LMW (>50) give the best score for ATENE (bottom left) and for NER performances (bottom right). This result is in line with those presented in [1] where the authors reported that a larger LMW gives a better speech translation score.

In order to check which ASR error metric best correlates with the NER error metric, we measured correlations between the results of the NER metric (ETER) and the results produced by the three ASR metrics (WER, NE-WER and ATENE). Figure 3 shows the linear regression of WER, NE-WER and ATENE vs ETER.

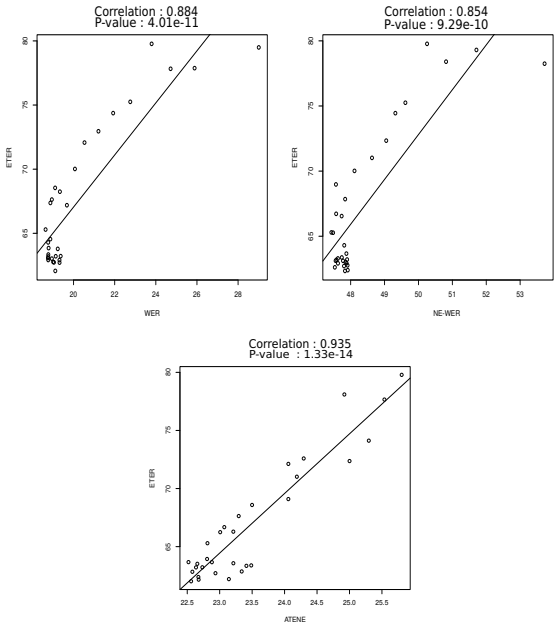

Figure 3: *Pearson correlation between NER error metric (ETER) and ASR error metrics (WER, NER-WER and ATENE)*

As we can see, ATENE (0.93) correlates better than WER (0.88) and NE-WER (0.85) with the actual NER system performance. ATENE can thus be considered a better predictor of the NER results than the other two metrics.

4.3.2. End-to-end results

To assess the full impact of the ATENE-based tuning regimen, we selected the optimal tuning parameter values chosen by each metric, and then run and evaluated the NER system on the resulting hypotheses. In addition, we also ran a tuning step with the NER system in the loop to estimate what the best possible result could be, which we call the "oracle" run. We followed the same data/test split methodology as in the previous experiment. The results shown correspond to the mean of the 21 individual results. Results are shown table 2.

As shown in Table 2, the optimization of the LM weight and word insertion penalty using the NER system performance directly leads to an absolute 1.09% ETER improvement as compared to the optimization based on WER. Therefore if the NER system of the final application is available for ASR developers, it will be beneficial to use it directly to tune their system. However, if the NER system is not available, using ATENE can be considered a good alternative to WER since it provides an absolute gain of 0.78% of ETER. By checking the value of the ASR feature weights, we found that a relatively large LM weight value is obtained both when using ATENE (56) or the NER performance (58) as a tuning metric. A possible explanation is that by putting a higher weight on the language model, the ASR will generate more grammatically structured sentences. This may be preferred by the NER system, hence a better ETER score, despite an increase in WER.

Table 2: *NER performance comparison on ASR tuned according to WER, NE-WER, ATENE and using ETER (oracle).*

Metrics used for tuning ASR	WER	NER error rate
WER	16,92	64.90%
NE-WER	17,05	65.33%
ATENE	17,24	**64.12%**
NER (oracle)	17,37	**63.81%**

5. Conclusion and future work

In this paper we investigated different ways to tune ASR systems for a speech-based NER system. We tested several ASR metrics, including WER, NE-WER and ATENE. Our experiments were run in two different perspectives. First, the NER system was unknown during system development; second, the NER system was known (oracle). Results show that when using a NER oracle for ASR tuning, the NER error rate drops by more than 1% absolute, while the first configuration achieves at best a 0.75% gain with the ATENE metric.

We first studied the impact of the tuning parameter changes on different ASR metrics such as WER, NE-WER, ATENE. This leads to the observation that the optimal language model weight changes with the metrics. In addition, and in line with the conclusion presented in [1] in the speech translation field, we observed that the optimal NER performance is reached with a higher LM weight than a WER-based optimization provides. We thus studied the correlation between these ASR metrics and the NER performance, and reached the conclusion that within the tested metrics ATENE is providing the best prediction for NER performance. Secondly, we measured the final NER performance on the ASR system outputs corresponding to the different optimization points provided by the ASR metrics, and compared the NER score with the best possible value reachable with the NER system in the optimization loop. This experiment has shown that WER is not optimal, and that optimizing on the ATENE metric allows to recover 70% of the performance loss compared to the maximum possible gain (oracle).

The optimisation experiment presented in this work is extremely simple and already yields an interesting gain while adjusting only a low number of parameters (language model weight and word penalty) during the tuning procedure. This shows that some ASR tuning can be done without explicitly knowing the subsequent NER system. We also show that the proposed optimization entails a higher ASR language model weight which results in better NER performance, despite a local increase of WER.

As a future work we aim to explore more ASR parameters including the decoding algorithm, number of Gaussians or features used in the HMM. We also think that the NER system should be adjusted to better deal with ASR errors that we cannot eliminated, which will require a better understanding of the impact of ASR errors on the NER system.

6. References

[1] 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.

[2] 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'99 Proceedings*. Morgan Kaufmann Pub, 1999, p. 215.

[3] M. A. Ben Jannet, O. Galibert, M. Adda-Decker, and S. Rosset, "How to evaluate asr output for named entity recognition?" in *16th Annual Conference of the International Speech Communication Association (Interspeech'15)*, 2015.

[4] 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.

[5] P. R. Dixon, A. Finch, C. Hori, and K. Hideki, "Investigation on the effects of asr tuning on speech translation performance," in *IWSLT*, 2011.

[6] 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.

[7] A. L. Berger, V. J. D. Pietra, and S. A. D. Pietra, "A maximum entropy approach to natural language processing," *Computational linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[8] 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)*. Istanbul, Turkey: European Language Resources Association (ELRA), may 2012.

[9] O. Galibert, S. Rosset, C. Grouin, P. Zweigenbaum, and L. Quintard, "Structured and extended named entity evaluation in automatic speech transcriptions." in *IJCNLP*, 2011, pp. 518–526.

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

[11] M. Marrero, J. Urbano, S. Sánchez-Cuadrado, J. Morato, and J. M. Gómez-Berbís, "Named entity recognition: Fallacies, challenges and opportunities," *Computer Standards & Interfaces*, pp. 482 – 489, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0920548912001080

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

[13] 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 Fifth Linguistic Annotation Workshop (LAW-V)*. Portland, OR: Association for Computational Linguistics, June 2011, pp. 92–100.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] J. Gauvain, L. Lamel, and G. Adda, "The limsi broadcast news transcription system," *Speech Communication*, vol. 37, pp. 89–108, 2002.

[20] L. Lamel and J.-L. Gauvain, "Speech processing for audio indexing," in *Advances in Natural Language Processing*. Springer Berlin Heidelberg, 2008, pp. 4–15.

[21] 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

[22] M. B. Jannet, M. Adda-Decker, O. Galibert, J. Kahn, and S. Rosset, "Eter: a new metric for the evaluation of hierarchical named entity recognition," in *LREC'14*, Reykjavik, Iceland, may 2014.