Reshaping Automatic Speech Transcripts for Robust High-level Spoken Document Analysis

Julien Fayolle
INRIA
IRISA, Rennes - FRANCE
julien.fayolle@irisa.fr

Fabienne Moreau
University Rennes 2
IRISA, Rennes - FRANCE
fabienne.moreau@irisa.fr

Christian Raymond
INSA
IRISA, Rennes - FRANCE
christian.raymond@irisa.fr

Guillaume Gravier
CNRS
IRISA, Rennes - FRANCE
guillaume.gravier@irisa.fr

ABSTRACT

High-level spoken document analysis is required in many applications seeking access to the semantic content of audio data, such as information retrieval, machine translation or automatic summarization. It is nevertheless a difficult task that is generally based on transcripts provided by an automatic speech recognition system. Unlike standard texts, transcripts belong to the category of highly noisy data because of word recognition errors that affect, in particular, very significant words such as named entities (e.g. person's names, locations, organizations). Transcripts also contain specificities of spoken language that make ineffective their processing by natural language processing tools designed for texts. To overcome these issues, this paper proposes a method to reshape automatic speech transcripts for robust high-level spoken document analysis. The method consists in conceiving a new word-level confidence measure that may efficiently ensure the reliability of transcribed words, focusing on words that are relevant for high-level spoken document analysis such as named entities. The approach consists in combining different features collected from various sources of knowledge thanks to a machine learning method based on conditional random fields. In addition to standard features (morphosyntactic, linguistic and phonetic), we introduce new semantic features based on the decisions of three robust named entity recognition systems to better estimate the reliability of named entities. Experiments, conducted on the French broadcast news corpus ESTER, demonstrate the added-value of the proposed word-level confidence measure for error detection and named entity recognition, with respect to the basic confidence measure provided by an automatic speech recognition system.

AND'10, October 26, 2010, Toronto, Ontario, Canada.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*language parsing and understanding, speech recognition and synthesis*

Keywords

Spoken Language Processing, Automatic Speech Recognition System, Noisy Transcription, Word-level Confidence Measures, Feature Combination, Named Entities, Machine Learning

1. INTRODUCTION

With the advent of digital television, the many channels that have emerged and the proliferation of podcasts or video webcasts (e.g. video on YouTube), the amount of multimedia information—sound, speech, image, text and video—produced in recent years continues to grow. As with the profusion of web pages and text content a few years ago, high-level applications able to efficiently handle this amount of audiovisual data are required. Users need relevant information retrieval systems and, for instance, automatic tools for summarization, translation or topic tracking of multimedia content. Nevertheless, these applications are more challenging to conceive for multimedia data than for text documents. Indeed, they require an understanding of multimedia content, which means being able to extract semantic knowledge. For this purpose, and considering that the speech remains the best—or at least the easiest—way to extract semantic content from multimedia data, most applications primarily rely on the audio track of multimedia data, and more particularly on transcripts provided by automatic speech recognition (ASR) systems.

Unlike traditional text documents (i.e. written text), transcripts belong to the category of highly noisy data and are complex to process for three reasons. Firstly, transcripts are raw, i.e. they contains no sentence segmentation, no punctuation mark, and, in some cases, no capitalization (e.g. uppercase). This is already a major issue because most of the existing natural language processing tools were initially designed for texts and tend to rely on punctuation and capital letters. Secondly, spoken language contains many particularities (e.g. hesitations, repetitions, corrections, grammar quite different from written language) which makes it difficult to process and require more suitable tools and tech-

niques. Lastly, the main issue arises from the inevitable recognition errors contained in automatic transcripts. Indeed, word error rate may vary from 10 % to over 60 % depending on the type of content considered. One cause of these errors is related to the inner working of ASR systems which are based on closed lists of vocabulary to transcribe speech. All out-of-vocabulary words are automatically replaced by other in-vocabulary words, acoustically close but erroneous. Among the out-of-vocabulary words, some words such as named entities (NEs i.e. person's names, locations, organizations, etc.) are highly meaningful and relevant for many tasks related to spoken language processing (SLP). For instance, in the field of information retrieval, a search engine log analysis shows that a significant proportion of user queries are NEs. If these words are misrecognized in automatic transcripts, they cannot be matched with query words and, as a consequence, the relevant information cannot be retrieved.

In the absence of appropriate tools for the automatic extraction of linguistic (or semantic) information from transcripts, most studies intending to exploit transcripts for SLP tasks work at the word-level. Although words in the transcripts may be erroneous, their exploitation in some applications may be sufficient. For instance, it has been shown that the application of a textual information retrieval system on automatic transcripts may give acceptable or similar results compared to those observed on written texts [8]. However, this observation may be mitigated since several work showed that these results are obtained only if transcripts contain a reasonable error rate (less than 40 %) and in very favorable experimental conditions (e.g. long queries) [4]. This conclusion is problematic since in many cases, especially when working on very noisy data such as TV data, transcripts might exhibit word error rates higher than 40 %.

One key to the problem would be to exploit only words that were correctly recognized by the ASR system. Nevertheless, this solution is possible only if measures indicating the reliability or unreliability of words are available. Many works focused on this problem and have shown that confidence measures (CMs) provided by ASR systems are not effective enough to correctly distinguish recognized words from recognition errors [16]. Moreover, in many applications, words are not sufficient to describe the semantics of the data. Linguistic information, such as semantic knowledge, are therefore required in order to carry out more elaborate processings. For automatic summarization or machine translation tasks for instance, the benefit of using semantic information such as NEs is undeniable as clearly demonstrated in [1, 11]. The main issue is therefore to develop new techniques able to extract semantic information and sufficiently robust to be applied on noisy data such as automatic speech transcripts.

In this context, the paper proposes a method to reshape automatic speech transcripts for robust high-level spoken document analysis. This method consists in building a new word-level CM able to more efficiently ensure the reliability of the transcribed words, focusing in particular on the words such as NEs that are useful for high-level SLP tasks. The approach consists in combining different features collected from various sources of knowledge by means of a machine learning method based on conditional random fields. In addition to the standard morpho-syntactic, linguistic and phonetic features, we introduce new semantic features based on the decisions made by three robust named entity recognition (NER) systems to better estimate the reliability of named entities.

The paper is organized as follows. Section 2 introduces the context of this study and related work. Section 3 details the proposed approach to obtain a new word-level CM, more reliable for high-level SLP applications. Section 4 describes the experimental setup before reporting experiments and results regarding the contribution of CMs for error detection and for NER. Finally, conclusions are given in Section 5.

2. STATE-OF-THE-ART

Automatic speech recognition (ASR) is a complex task and, in spite of recent progress, state-of-the-art systems still produce automatic transcripts that may contain many recognition errors. This is particularly true when ASR systems are applied on radio or TV streams with unprepared spontaneous speech or with background noise. Most of the semantic content of these digital documents is contained in speech and thus high level tasks (summarization, information retrieval, machine translation, etc.) need to exploit automatic transcripts given by an ASR system. The presence of errors in automatic transcripts is the main difficulty and, to improve robustness, confidence measures (CMs) have been widely investigated.

Among existing work on CMs improvement, some propose to estimate the confidence of a word, directly as its a posteriori probability, given low-level (acoustic) observations [14, 16]. These methods provide fairly good results, especially when the probabilities are estimated from N-best hypothesis lists or word graphs [27], but they are dependent on a given ASR system and involve modifying its inner workings. Moreover, better CMs can be designed by exploiting sources of knowledge in addition to the ASR system's resources. Another way to compute CMs is then to search for relevant additional clues, within the ASR output, that are sufficiently informative to distinguish correctly recognized words from possible recognition errors. These clues, called (predictor) features, are generally obtained during the decoding phase at either the acoustic or the language model level or from other sources of knowledge (syntactic, semantic, etc.). To improve performance, these features are combined together, potentially including the low-level confidence scores provided by the ASR system, and transformed into a single CM that indicates the reliability of the recognized words. Many confidence features have been studied [16] but the solutions proposed for word-level CMs for large vocabulary speech recognition are limited to combine low-level information (strongly dependent on the ASR system). Higher level of information independent from the ASR system have been investigated, but restricted to the out-of-vocabulary problem [19] or small vocabulary tasks such as spoken dialogs [12, 25, 28].

This work aims at conceiving a word-level CMs useful to be effective for a wide range of spoken language processing tasks. The difficulty lies in the fact that the resulting word-level CMs should be relevant for all words and for transcripts generated by any ASR system. We propose to use different predictors, independent from the ASR system, in order to improve any decoder base measure. New high-level features designed from the decision of several robust named entity recognition systems are also introduced to particularly im-

position i :	1	2	3	4	5	6	7
relative position j :	-3	-2	-1	0	+1	+2	+3
sequence:	le	tour	de	france	troisième	étape	remportée
X^{lmbb}**:**	11	22	33	I4	I3	I2	I3
X^{feat}**:**	$X^{feat}_{i=1,j=-3}$	$X^{feat}_{i=2,j=-2}$	$X^{feat}_{i=3,j=-1}$	$X^{feat}_{i=4,j=0}$	$X^{feat}_{i=5,j=+1}$	$X^{feat}_{i=6,j=+2}$	$X^{feat}_{i=7,j=+3}$
$X^{\cdots}$**:**	...	...	...	...	...	...	...
Y**:**	correct	correct	correct	correct	correct	correct	correct
$CM_i = p(Y_i = correct\|X)$**:**	0.91	0.96	0.98	0.96	0.97	0.90	0.84

Table 1: Example of the current word 'france' in its context where X represents the sequence of features used to estimate the label sequence Y and the confidence measure (i.e. marginal probability $p(Y_i = correct|X)$) for each position i.

prove word-level CMs on words pertaining to a named entity. This new CM is detailed in the following section.

3. ENRICHED WORD-LEVEL CONFIDENCE MEASURE

The automatic speech recognition (ASR) system used on this work provides a baseline confidence measure (CM) obtained from N-best sentence hypothesis lists as detailed in Section 3.1. However, we have experimentally observed that ASR-based CMs are not sufficiently accurate to ensure the reliability of transcribed words, and more particularly of semantically meaningful words. We propose to combine additionnal features to better estimate the reliability of meaningful words. In addition to the standard phonetic, morphosyntactic and linguistic features described in Section 3.2, we introduce, in Section 3.3, new high-level features based on the decisions of three named entity recognition (NER) systems particularly robust to transcription errors. Moreover, we investigate the use of contextual features to improve CMs. Section 3.4 describes how the context is used to build contextual features from selected base features while Section 3.5 presents the conditional random field approach used to combine base and contextual features so as to provide an enriched CM.

3.1 ASR confidence measure

The ASR CM is provided by the ASR system used in this work and is derived from N-best lists, using a posteriori sentence probabilities obtained by the combination of an acoustic score, a linguistic score provided by a 4-gram language model (LM) and a morpho-syntactic score given by a 7-gram part-of-speech (POS) model [14]. The CM at the word level is given by summing the sentence a posteriori probability over all transcription hypotheses in the N-best lists in which the word appears in the correct position.

This ASR CM will be used as a baseline in the experiments. Moreover, such measure can also be used as a feature to be combined with the features described below. It is however important to note that any CM provided by any ASR system could have been used.

3.2 Standard features

One of our major requirement is that features should be as independent as possible from a particular ASR system and should be easily accessible, e.g. by post-processing the output information provided by the ASR system. A few standard features meeting these requirements were selected from previous work [6], covering three knowledge sources, namely morpho-syntactic, linguistic and phonetic.

Part-of-speech categories (**pos**) are used as morpho-syntactic features. Transcripts are tagged with a set of 144 POS classes containing general morpho-syntactic classes as well as very frequent words [14]. This feature enables to know the a priori error distribution for each POS class, assuming some classes are more error-prone than others.

The language model back-off behavior (**lmbb**) has proven to be a valuable error predictor [19, 22]. Given a language model and a word sequence, the language model back-off behavior indicates for each word the degree of back-off used in the LM, i.e. the degree n of the largest current n-gram belonging to the LM. The ASR 4-gram language model is used here for sake of simplicity but any LM could have been chosen, independently from the ASR system. The **lmbb** feature is composed of 4 main classes ('I1', 'I2', 'I3', 'I4') and 6 other specific classes ('11', '21', '22', '31', '32', '33'), the latter representing the different cases at the beginning of a sequence. For a class xy, x and y represent respectively the position in the sequence (the first 3 positions '1', '2', '3' and the next ones 'I' inside the sequence) and the LM degree n as defined previously (n from 1 to 4). An example is given in Table 1.

Finally, the number of phonemes (**#ph**) and the total word's duration (**dur**) are used as phonetic features. Indeed, many observations points out that word length can help in predicting correct words and errors. For example, out-of-vocabulary words tend to be misrecognized as a sequence of short words while long in-vocabulary words in the ASR output are often correctly hypothesized.

3.3 High-level features based on named entity recognition

In order to obtain a CM that is particularly efficient on semantically meaningful words (e.g. NEs), we propose to use, in addition to the base predictors previously presented, higher-level features based on NER.

3.3.1 Benefit of named entities

Most spoken language processing (SLP) applications relying on automatic transcripts, like information retrieval, automatic summarization or machine translation, are generally based on lexical units without access to the explicit semantic content conveyed by words. Many past and present works try to discover semantic content in text to improve SLP systems. Unfortunately, methods are often restricted to a particular semantic domain, application dependent and moreover inefficient when applied to automatic speech transcripts.

Actually, NEs are pieces of semantic very generic which can be found in many documents and their usefulness to

system	HMM	SVM	CRF	oracle (SVM+CRF)	oracle (HMM+SVM+CRF)	best Ester 2 system on manual transcripts
manual transcript	27.89	28.06	22.79	/	/	9.80
automatic transcript	59.44	59.83	53.49	50.40	45.80	66.22

Table 2: Slot error rate [20] performance for the three NER systems. Results are also reported for the Oracle combination of systems and for the system that best performed on reference transcripts during the ESTER 2 campaign.

improve the quality of natural language processing systems have been demonstrated on several occasions. For instance, [1, 26] used NEs to improve machine translation and [11] for automatic summarization. NEs are also crucial for precise information retrieval systems such as Question-Answering systems [17]. Assuming that it is also crucial for most high-level SLP tasks to correctly recognize NEs in transcripts, we propose a robust method for NE detection whose results will be helpful to improve the CMs on NE words.

3.3.2 Robust named entity recognition systems

NER has been investigated via many evaluation campaigns, such as "Message Understanding Conference", "Automatic Content Extraction", "Document Understanding Conference", or "ESTER" for the French language. In the last French campaign "ESTER 2", the best systems are able to recognize more than 90 % of the NEs in the reference, human generated, transcripts. These systems are based on formal language description (e.g. [2]) but are not robust to process noisy automatic transcripts in which only 35 % of the NEs are correctly transcribed. In previous work [24], we have proposed to exploit together three different NER systems which have been proved to be robust for processing noisy automatic transcripts. Each of them got honorable performances on the ESTER 2 test data, achieving slightly better performance than those reported in the evaluation's final results for the most difficult condition [7] (ASR with an error rate around 26 %).

Three different systems have been investigated for the two reasons. Firstly, because of the lack of robustness one system is not sufficient, so we want to investigate several ones to take advantages of their own specificities to process automatic transcripts. Secondly, the chosen machine learning algorithms must differ each other significantly to offer the most orthogonal decisions as possible and then to better estimate the reliability the recognized NEs.

The following machine learning algorithms have been chosen to design the three systems:

- *Conditional Random Field* (CRF): this algorithm has been successfully applied to NER [23] and seems to be one of the most efficient algorithm in the case of sequence labeling problem [10];
- *Support Vector Machines* (SVM): this algorithm has been intensively investigated for NER (e.g. [15]) and subsequently differs from CRF-based approaches;
- *Hidden Markov Model* (HMM): discriminant algorithms such as CRF or SVM tends to outperform generative ones with a sufficient amount of training data but are difficult combine with ASR systems. For instance, [21] emphasizes that improving only NE models does not yield significant improvements and therefore suggests better integration between NER and ASR. The same conclusion is drawn in [9]: state-of-the-art approaches only post-process the best transcription hypothesis and performance will therefore always be correlated with word error rates. To bridge the gap between ASR and NER, several authors investigated the use of confusion networks containing multiple transcription hypotheses [3, 5, 13]. For these reasons, a HMM process, represented as finite state transducers, has been developed as a third NER system to be able to efficiently process word graphs later.

All three systems, detailed in [24], were evaluated on the French ESTER 2 data on the following categories: person's names, functions, organizations, localizations, human productions (movies, books), dates and times, and amounts. Performance are summarized in Table 2 and Oracle results for system combination are reported. Oracle results show clearly that the system decisions are complementary since the potential of NER improvement is about 7 points error reduction (45.8 % against 53.5 %). This complementarity will be exploited in the next section as semantic feature for word error CM.

3.3.3 High-level named entity features

From the decisions of the three NER systems, we propose to build five high-level features which are the following:

1. the NE category recognized by the CRF system;
2. the probability of the recognized NE category given by the CRF system;
3. the NE category recognized by the SVM system;
4. the NE category recognized by the HMM system;
5. the agreement between the 3 systems composed of 3 classes: 'NE' (if the 3 taggers recognize the same NE), 'NaNE' (if the 3 taggers recognize that the word is Not a NE), '?' (in ambiguous cases). We believe that these ambiguous cases are induced by recognition errors, especially on semantically rich words.

The specificity of our approach is to propose a real interaction between the SLP techniques and the CMs. On the one hand, we believe that word-level CMs are useful for SLP applications to detect the presence of recognition errors in the transcripts. On the other hand, we believe that SLP techniques (such as NER) may also, by feedback, be useful to better estimate the confidence scores. We can also note that apart from NEs, other semantic features could be designed in the same way using different SLP applications to identify semantic ambiguities.

3.4 Contextual features

It is a well-known fact that a transcription error often impacts the surrounding words and the use of context have

already shown promising results in error detection in different contexts [19, 25]. Thus, we propose to add contextual features from each of the base feature described previously to better estimate the CM [6]. The process includes two steps: defining the context and chosing the relative positions in the sequence to add as new contextual features. Firstly, the *context* at a current position in a sequence is composed of the s neighbors on both sides. Empirically, we chose $s = 3$ as we observed that a longer context size was not necessary. Table 1 gives an example of the word 'France' in its context. Secondly, we chose all the *relative positions* included in the context to create new features. For example, in table 1, all the features in the 6 relative positions '-3', '-2', '-1', '1', '2' and '3' are used as contextual features in addition to the current features in position '0'. This means that for each current position (e.g. $i = 4, j = 0$) and for each kind of feature (e.g. the lmbb feature), we use 1 current feature (e.g. $X^{lmbb}_{i=4,j=0}$ = I4) and 6 contextual features (e.g. $X^{lmbb}_{i=1,j=-3} = 11$, $X^{lmbb}_{i=2,j=-2} = 22$, $X^{lmbb}_{i=3,j=-1} = 33$, $X^{lmbb}_{i=5,j=+1}$ = I3, $X^{lmbb}_{i=6,j=+2}$ = I2, $X^{lmbb}_{i=7,j=+3}$ = I3).

3.5 CRF combination

As described in Section 2, a classical approach to compute CMs is to combine relevant features that are sufficiently informative to indicate the reliability of the recognized words. Many combination models (SVM, boosting, decision trees, hidden Markov models, conditional random fields, etc.) have already been investigated in literature [12, 16, 28]. In the context of this work, we use a machine learning method based on conditional random fields (CRF). This choice is motivated by several reasons. Firstly, we need a machine learning algorithm that is flexible enough to deal with features collected from different sources of knowledge. The CRF model, which is by construction a discriminant model, is thus well suited to accommodate many statistically correlated features as input. Secondly, a CRF is a probabilistic model especially dedicated to labeling sequential data [18]. Since it is clear that in transcripts the word recognition errors are highly dependent one on another, it is important to have an algorithm able to manage sequential data for labeling automatic transcribed words with the labels 'correct' or 'erroneous'. So, unlike many other discriminant models (SVM, perceptron, etc.) that view the sequential labeling problem as a set of independent decisions, CRFs compute the conditional probability $p(Y|X)$ of a sequence of labels Y given a sequence of observations X, thus taking a global decision on the sequence. Moreover, CRFs are able to easily estimate a marginal probability $p(Y_i = y|X)$ of each decision at the position i in the sequence, the marginal probability of the label 'correct' being used as the word-level CM. All the details on CRFs for segmenting and labeling sequence data can be found in [18]. Finally, another benefit of the CRF classifier is its ability to weight each feature during the training stage, making it possible to interpret the most relevant "rules" for predicting correct or erroneous words.

Since CRFs[1] consider symbolic features, all continuous features previously presented were discretized using a C4.5 decision tree so as to minimize the entropy of each class.

4. EXPERIMENTS

[1]CRF++ (http://crfpp.sourceforge.net/) is used in this work.

In this section, three experiments are carried out in order to evaluate the performances of four confidence measures ('ASR CM', 'ASR+S CM', 'ASR+NE' and 'ASR+S+NE') combining gradually the ASR CM ('ASR'), the standard features ('S') and the named entity features ('NE'). These experiments aim at evaluating the benefit of using CMs: firstly, for error detection on all transcribed words (Section 4.2.1), recognized NEs (Section 4.2.2) and reference NEs (Section 4.2.3); secondly on data containing various word error rates (Section 4.3); and lastly, for a named entity recognition (NER) task (Section 4.4).

4.1 Experimental setup

Experiments are carried out with a large vocabulary radio broadcast transcription system, exhibiting error rates around 20 % on broadcast news data. Hidden Markov phonetic models were trained using approximately 200h of speech material. A 4-gram language model was obtained from about 500 million words mostly coming from French-speaking newspapers. The CM is provided based on posterior probabilities combining acoustic, language model and part-of-speech scores as in [14]. Results are reported on the corpus from the French evaluation campaign ESTER 2 [7], consisting of 12 hours of different French radio broadcasts for which NEs are manually annotated and for which word error rates from 16.0 % to 42.2 % were achieved. We used a 5-fold cross-validation on the whole corpus to evaluate performances: 80 % for training sets and 20 % for test sets. The 5 folds were built in order to distribute uniformly each radio broadcast in all folds. The results were obtained by averaging the five rounds of cross-validation. In the experiments, we evaluate the CMs focusing on 3 different sets obtained by filtering the words in the output of the CRF classifier: the first set contains all transcribed words of the transcripts (no filtering), the second one all the NEs recognized by the NER systems, and the last one all the reference NEs that was annotated manually. The standard evaluation metric of equal error rate (EER) and the receiver operating characteristic (ROC) curve are used to evaluate performances for error detection. The precision/recall curve is used to evaluate performances for NER. The revised word error rate (RWER), that represents the sum of insertion and substitution rates, is used to evaluate the quality of transcripts.

4.2 Error detection

In this experiment, we evaluate the benefit of using CMs for error detection on all transcribed words, recognized NEs and reference NEs. General results are reported in Table 3.

4.2.1 On all transcribed words

The ROC curve on Figure 1 shows the performances of the different measures for all transcribed words. All the combinations exhibit better performances than the baseline (ASR CM), the one with the standard features 'S' shows the best improvement in EER (24.61 vs. 29.21). The NE features allow a significant improvement at high false error detection rate when combined with the ASR CM. Nevertheless, these features allow a very small improvement of the performance when combined with the standard features. This can be explained by the fact that NEs represent only 18.95 % of the whole corpus.

4.2.2 On all recognized named entities

	all	rec NEs	ref NEs
# words	118891	19642	22529
% corpus	100.00	16.52	18.95
RWER	23.02	17.66	23.46
ASR CM	29.21	29.69	26.87
ASR+NE CM	29.11	26.88	24.46
ASR+S CM	24.61	23.07	20.25
ASR+S+NE CM	**24.37**	**22.22**	**19.36**

Table 3: Results for each confidence measure (CM) on all transcribed words (all), recognized named entities (rec NEs) and reference named entities (ref NEs) in number of words, revised word error rate (RWER) and equal error rate (EER).

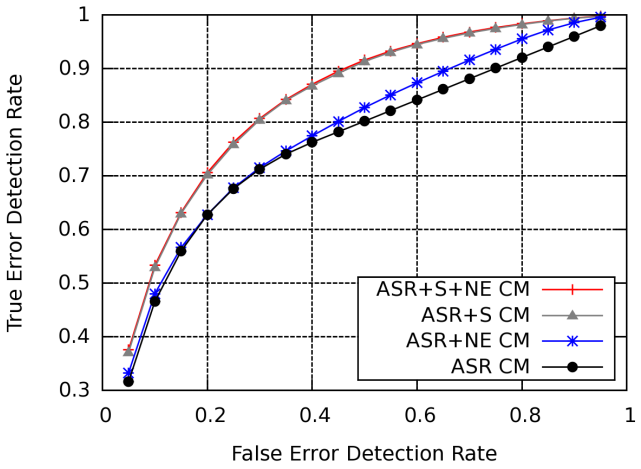

Figure 1: ROC curve showing the performances of the different confidence measures for *all transcribed words*.

The ROC curve on Figure 2 shows the performances of the different measures for all recognized NEs which represent 16.52 % of all transcribed words. A word is recognized as a part of a NE if it is recognized by at least one of the three NER systems. All the combinations exhibit better performances than the baseline (ASR CM), the one with the standard features shows again the best improvement in EER (23.07 vs. 29.69). The combination with the NE features has also significant results (26.88 vs. 29.69), this shows that using three NER systems can improve the CM on recognized NE. Nevertheless, the benefit of all features is not the sum of the individual improvement (22.22 vs. 29.69) and using NE features in addition of standard features allow a limited improvement from 23.07 to 22.22 in EER. This means that the information conveyed by the NE features are partially redundant with the standard features but that they also contain new information useful to improve the CM.

4.2.3 On all reference named entities

Reference NEs (i.e. NEs that are manually tagged), are used in this experiment to verify the added-value of our approach in the case that we would have an ideal NER system able to recognize all NEs in the transcripts. The ROC curve on Figure 3 shows the performances of the different measures for all reference NEs. This experiment shows the potential gain of the proposed CMs on perfectly recognized NEs. The

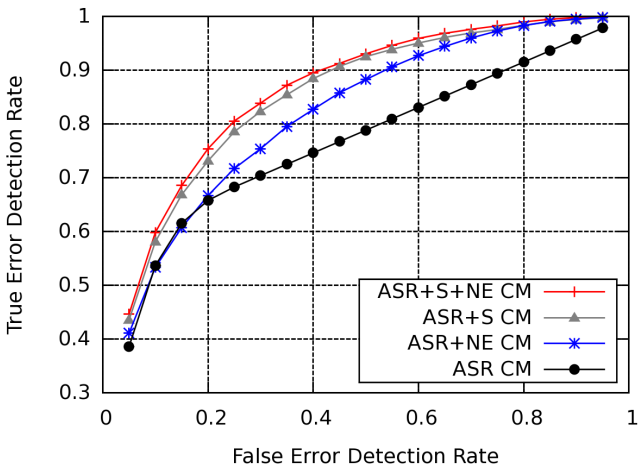

Figure 2: ROC curve showing the performances of the different confidence measures for *all recognized named entities*.

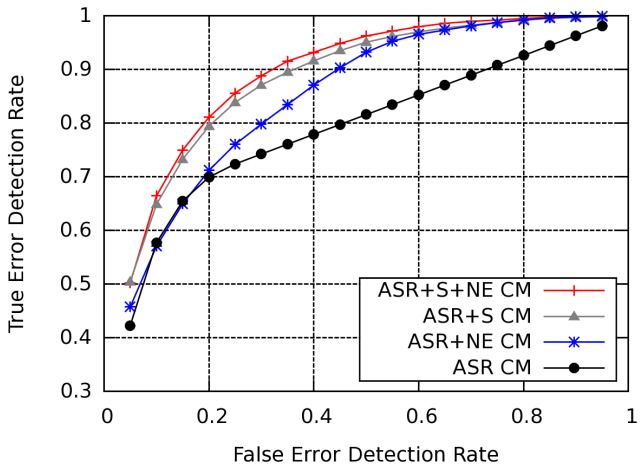

Figure 3: ROC curve showing the performances of the different confidence measures for *all named entities of reference*.

curves have the same aspect than in the previous experiment but with better gains.

4.3 Noise influence

In this experiment, the influence of noise (estimated in term of revised word error rate) on the performances of the confidence measure for the error detection task on all transcribed words, recognized NEs and reference NEs is evaluated. Results reported in Table 4 show that: firstly, for all kinds of radio broadcast (with revisited word error rate ranging from 16% to 29%), the proposed CMs are better than the baseline; secondly, in some cases, the NE features bring a slight improvement in addition of the ASR+S features.

4.4 Enriched confidence measure for a named entity recognition task

In this experiment, the impact of using our CMs for a SLP task such as named entity recognition (NER) is evaluated. For each NE recognized by the three NER systems,

radio:	inter	rfi	tvme	africa1
All transcribed words				
RWER	21.85	16.18	20.45	29.02
ASR CM	31.63	25.40	26.50	28.57
ASR+NE CM	30.94	23.82	25.55	28.55
ASR+S CM	27.27	**20.18**	**20.29**	23.78
ASR+S+NE CM	**26.93**	20.48	20.39	**23.58**
NE recognized				
RWER	18.77	12.32	16.93	19.84
ASR CM	31.90	28.93	25.92	28.83
ASR+NE CM	29.07	25.09	22.54	26.31
ASR+S CM	24.82	20.45	19.61	22.72
ASR+S+NE CM	**24.62**	**20.13**	**17.94**	**22.24**
NE reference				
RWER	16.92	16.82	24.43	31.39
ASR CM	28.62	24.71	25.23	27.00
ASR+NE CM	27.03	21.04	20.74	24.58
ASR+S CM	23.35	18.80	16.34	20.47
ASR+S+NE CM	**22.42**	**18.47**	**15.51**	**19.50**

Table 4: Results for each confidence measure (CM) on specific parts of the corpus with different revised word error rate (RWER) in equal error rate.

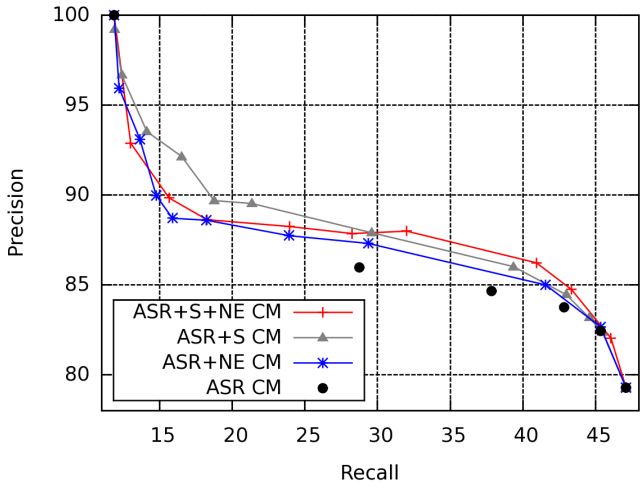

Figure 4: Precision/recall curve showing the performances of named entity recognition using the different confidence measures.

the mean of its confidence measures is computed and compared to a given threshold to decide to keep or not to keep the recognized NE. Varying the threshold from 0 to 1, a precision/recall curve is obtained (Figure 4) for each evaluated CM. The observed results for the precision rate show that our enriched CM is especially effective to filter misrecognized NE. Therefore, the main benefit of this CM for the SLP task is that they can be used to efficiently ensure the reliability (or the unreliability) of some meaningful words such as NEs.

5. CONCLUSIONS

This work tackles the difficult problem of exploiting noisy data such as automatic speech transcripts for high-level spoken document analysis. A conditional-random-field-based combination of standard and high-level features have been proposed to improve word-level confidence measures ensuring the reliability of the transcribed words, focusing especially on the words such as named entities that are meaningful for high-level spoken language processing (SLP) tasks. Experiments conducted on a french radio broadcast corpus show several interesting results. Firstly, we have confirmed that our feature combination is beneficial to improve the confidence measure provided by an automatic speech recognition (ASR) system. A significant improvement is obtained to detect errors on all transcribed words but also to recognize erroneous words more specifically related to named entities (recognized by our system or by an ideal system). These results are confirmed on several different kind of corpus containing various word error rates. Secondly, the selected features can be obtained independently of any ASR system and directly from the output provided by systems or external tools such as language processing ones. This point is crucial in the context of this work whose main focus is to obtain more reliable confidence measures for use by high-level SLP techniques independently from a particular ASR system. Thirdly, experiments have also shown the benefit of introducing in the combination a high-level feature, obtained by a robust SLP system, to detect errors related to named entities. Nevertheless, the added-value of this feature compared to a standard feature combination remains rather limited. Several reasons may explain this weak improvement. First, the high-level features contain information that are certainly redundant with those of standard features. Further experiments should thus be conducted to explain this redundancy. The second reason is probably due to the difficulty we encountered in discretizing the continuous features so that it can be handled by our machine learning algorithm that only considers symbolic features. The development of a more effective method able to better manage the conversion of digital to symbolic data is also part of our ongoing work. Finally, based on the idea that a veritable interaction between SLP techniques and confidence measures is beneficial to improve access to the semantic content of noisy transcripts, future works will also aim at finding other high-level confidence features required for robust high-level spoken document analysis.

6. REFERENCES

[1] B. Babych and A. Hartley. Improving machine translation quality with automatic named entity recognition. In *Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tool*, pages 1–8, Morristown, NJ, USA, 2003.

[2] C. Brun and M. Ehrmann. Adaptation of a named entity recognition system for the ester 2 evaluation campaign. In *IEEE International Conference on Natural Language Processing and Knowledge Engineering*, Dalian, Chine, 2009.

[3] F. Béchet, A. Gorin, J. Wright, and D. Hakkani-Tur. Named entity extraction from spontaneous speech in *How May I Help You ?* In *Proceedings of the International Conference on Spoken Language Processing*, Denver, Colorado, USA, 2002.

[4] C. Chelba, T. J. Hazen, and M. Saraclar. Retrieval and browsing of spoken content. *Signal Processing Magazine, IEEE*, 25(3):39–49, 2008.

[5] B. Favre, F. Béchet, and P. Nocéra. Robust named entity extraction from spoken archives. In *Proceedings*

of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, pages 491–498, Morristown, NJ, USA, 2005.

[6] J. Fayolle, F. Moreau, C. Raymond, G. Gravier, and P. Gros. Crf-based combination of contextual features to improve a posteriori word-level confidence measures. In *International Conference on Speech Communication and Technologies, Interspeech'10*, Makuhari, Japan, 2010.

[7] S. Galliano, G. Gravier, and L. Chaubard. The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *International Conference on Speech Communication and Technologies, Interspeech'09*, pages 2583–2586, 2009.

[8] J. S. Garofolo, C. G. P. Auzanne, and E. M. Voorhees. The trec spoken document retrieval track: A success story. In *Text Retrieval Conference (TREC-8)*, pages 16–19, 2000.

[9] Y. Gotoh and S. Renals. Information extraction from broadcast news. *Philosophical Transactions of the Royal Society of London, Series A*, 358:1295–1310, 2000.

[10] S. Hahn, P. Lehnen, C. Raymond, and H. Ney. A comparison of various methods for concept tagging for spoken language understanding. In *Proceedings of the Language Resources and Evaluation Conference*, Marrakech, Morocco, May 2008.

[11] M. Hassel. Exploitation of Named Entities in Automatic Text Summarization for Swedish. In *Proceedings of NODALIDA'03 - 14th Nordic Conference on Computational Linguistics*, Reykjavik, Iceland, 2003.

[12] T. Hazen, S. Seneff, and J. Polifroni. Recognition confidence scoring and its use in speech understanding systems. *Computer Speech & Language*, 16(1):49–67, 2002.

[13] J. Horlock and S. King. Named entity extraction from word lattices. In *Proceedings of European Conference on Speech Communication and Technology*, Geneva, 2003.

[14] S. Huet, G. Gravier, and P. Sébillot. Morpho-syntactic post-processing of n-best lists for improved french automatic speech recognition. *Computer Speech and Language*, 12(4):663–684, 2010.

[15] H. Isozaki and H. Kazawa. Efficient support vector classifiers for named entity recognition. In *Proceedings of the 19th international conference on Computational linguistics*, pages 1–7, Morristown, NJ, USA, 2002.

[16] H. Jiang. Confidence measures for speech recognition: A survey. *Speech communication*, 45(4):455–470, 2005.

[17] M. A. Khalid, V. Jijkoun, and M. De Rijke. The impact of named entity normalization on information retrieval for question answering. In *Proceedings of the 30th European conference on Advances in information retrieval*, pages 705–710, Berlin, Heidelberg, 2008.

[18] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA, 2001.

[19] B. Lecouteux, G. Linarès, and B. Favre. Combined low level and high-level features for Out-Of-Vocabulary Word detection. In *International Conference on Speech Communication and Technologies, Interspeech'09*, 2009.

[20] J. Makhoul, F. Kubala, R. Schwartz, and R. Weischedel. Performance measures for information extraction. In *the DARPA Broadcast News Workshop*, pages 249–252, 1999.

[21] D. Martin. Named entity extraction from speech: Approach and results using the textpro system. In *Proceedings of the DARPA Broadcast News Workshop*, pages 51–54, 1999.

[22] J. Mauclair, Y. Estève, S. Petit-Renaud, and P. Deléglise. Automatic detection of well recognized words in automatic speech transcription. In *Proceedings of the Language Resources and Evaluation Conference*, 2006.

[23] A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *CoNLL-2003*, pages 188–191, 2003.

[24] C. Raymond and J. Fayolle. Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. In *Traitement Automatique des Langues Naturelles*, Montréal, Canada, 2010.

[25] G. Skantze and J. Edlund. Early error detection on word level. In *COST278 and ISCA Tutorial and Research Workshop on Robustness Issues in Conversational Interaction*, 2004.

[26] R. K. Srihari and E. Peterson. Named entity recognition for improving retrieval and translation of chinese documents. In *Proceedings of the 11th International Conference on Asian Digital Libraries*, pages 404–405, Berlin, Heidelberg, 2008.

[27] F. Wessel, R. Schlüter, K. Macherey, and H. Ney. Confidence measures for large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, 9(3):288–298, 2001.

[28] D. Yu and L. Deng. Semantic confidence calibration for spoken dialog applications. In *ICASSP*, pages 4450–4453, 2010.