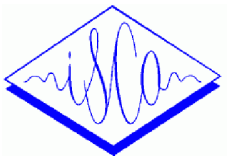
Automatic transcription error recovery for Person Name Recognition

Richard Dufour, Geraldine Damnati, Delphine Charlet, Frédéric Béchet



INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Automatic transcription error recovery for Person Name Recognition

Richard Dufour[1], *Géraldine Damnati*[1], *Delphine Charlet*[1], *Frédéric Béchet*[2]

[1]France Telecom - Orange Labs, Lannion, France
[2]Aix Marseille Université LIF-CNRS, Marseille, France
{richard.dufour,geraldine.damnati,delphine.charlet}@orange.com,frederic.bechet@lif.univ-mrs.fr

Abstract

Person Name Recognition from transcriptions of TV shows spoken content is a crucial step towards multimedia document indexing. Recognizing Person Names implies the combination of three main modules: Automatic Speech Recognition, Named-Entity Recognition and Entity Linking to associate the recognized surface form to a normalized Person Name. The three modules are potentially error prone. Hence, beyond each module's intrinsic complexity, the Person Names issue suffers from the highly dynamic evolution of vocabularies and occurrence contexts that are correlated to various dimensions (such as actuality, topic of the show…). This paper focuses on the first module and proposes an approach to recover from transcription errors made on Person Names. An error correction method is applied on the textual ASR output and we show that it is all the more efficient that it is coupled with a specific error region detection system. Experiments on the French REPERE database show that Person Names transcription can be efficiently corrected while preserving the overall transcription quality and thus increasing the performance of the whole Person Name Recognition process.

Index Terms: transcription error detection, transcription error correction, person name entity recognition.

1. Introduction

Person Name Recognition (PNR) is a particular case of Named Entity recognition which is particularly important from an applicative point of view. Indexing documents by the mentioned persons is relevant in itself but recognizing Person Names is also a preliminary step for further tasks such as Person Identification in documents. Person Identification is under the scope of the French REPERE[1] evaluation program for multi-modal person recognition in video documents. This program addresses both supervised and unsupervised identification, the latter requiring that no biometric model should be used and identification should be performed only by analysing the document itself. In this case, useful information regarding persons' identity can be found in the video through the detection and recognition of person names in overlaid texts and in the audio channel through the detection and recognition of person names in the spoken content. This paper addresses the second issue and proposes an enhanced process towards Person Name Recognition from automatic transcription of speech in TV contents.

A fully automatic PNR system is evaluated consisting of three modules: Automatic Speech recognition (ASR), Named Entity Recognition (NER) and Person Entity Linking (PEL). From a sequence of transcribed words "*la campagne de Hollande* (Hollande's campaign)", the NER module is in charge of localizing "*Hollande*" and assessing that it is a Person Entity (and not the French word for Netherlands) while the latter module allows the localized Person Entity to be associated to a normalized reference form (here *François_HOLLANDE*).

The three modules are potentially error prone. Hence, beyond each module's intrinsic complexity, the Person Names issue suffers from the highly dynamic evolution of vocabularies and occurrence contexts that are correlated to various dimensions (such as actuality, topic of the show, period when processing archives…). Even for the most efficient Speech-to-Text engines, the transcription of PN is by essence problematic. Some approaches have been designed to handle specific spoken term detection using sophisticated vocabulary adaptation techniques [1] or post-processing in multiple outputs (word lattices or sub-word unit lattices for hybrid approaches [2]). From a large scale applicative point of view, specifically tuning a system for a given content is not always feasible. In this perspective, we propose an alternative approach consisting of post-processing textual output provided by a state-of-the art industrial generic Speech to Text engine in order to recover from potential errors on PN.

The objective is to improve PN transcription without degrading the overall transcription quality. In fact, higher level Spoken Language Understanding processes (e.g. prediction of mentioned persons' presence in the show [3], named identification of speakers [4]) need to analyse the context in which PN occur. Hence, in order to exploit PNR it is essential to maintain high-quality transcription for the surrounding words.

Section 2 gives an overview of our Person Name Recognition system, presenting each of the three above mentioned modules. Section 3 presents the transcription error recovery approach and evaluations are provided in section 4.

2. Person Name Recognition system

2.1. Automatic Speech Recognition

Automatic transcription is performed with an industrial Speech to Text (STT) engine: the VoxSigma speech recognizer V3.5 from Vocapia Research, based on LIMSI technology [5]. This is a generic engine dedicated to Broadcast contents. This software provides speaker segmentation and speech recognition along with word-level confidence measures based on posterior probability for each recognized word. For genericity purpose and due to applicative constraints, we have chosen to directly exploit the one-best automatic transcription provided by this on-the-shelf STT engine. Hence we propose to improve the transcription of Person Names by only relying on the textual form of the STT one-best output, without any re-decoding process. The error recovery approach is detailed in section 3.

[1] funded by the ANR agency, http://www.defi-repere.fr

2.2. Named Entity Recognition

The Named Entity Recognition (NER) system used in this study is LIA_NE [6]. This system is based on a machine learning approach that includes two components: a generative HMM-based process used to predict Part-Of-Speech tags as well as semantic labels for each input word; a discriminative CRF-based process that takes as features the words as well as the syntactic and semantic labels given by the HMM in order to determine the span and the type of each entity detected.

The main advantage of a machine learning approach for NER when dealing with speech input is the ability to train the statistical model on a corpus that matches the output of ASR systems (no punctuation and unreliable capital letters). In LIA_NE the HMM and CRF models have been trained on the ESTER Broadcast News training corpus from which all punctuation and capitalization have been removed. This leads to increase the robustness of the system towards ASR errors [6].

2.3. Person Entity Linking

The goal of the Person Entity Linking module is to associate a unique identifier to each PN detected in a text, regardless of the way the name is expressed. For example, in the news corpus used to build this module, the current wife of the French president Sarkozy can be found under 6 forms: *Carla Bruni-Sarkozy, Carla Bruni, Mme Bruni-Sarkozy, Carla Sarkozy, Carla Bruni Sarkozy, Mme Sarkozy*. The PEL module is in charge of translating each of these forms into the normalized identifier: *Carla_BRUNI-SARKOZY*. Previous works on entity linking have focused on solving the linking ambiguities by using the context of the name occurrence in the document, either with heuristics or machine learning approaches [7]. We are dealing here with the opposite problem: given a potential list of persons that are likely to occur in TV shows, we want to predict all the different forms that can be used to express their names.

In usual frameworks, a closed list is chosen so as to be consistent with the ASR lexicon. Since we propose an approach to correct person names, we are not limited by the ASR lexicon. Hence, we have relaxed this constraint by considering a very large dictionary of person names, much larger that the one used by the ASR decoder. Each person name in this dictionary is associated to a normalized form. When there is an ambiguity (for example: *Mme Sarkozy* can refer to several persons) the most frequent normalized form is chosen. We built the person name dictionary used in this study with the following process:

- a corpus of newswire collected over the period 2004-2011: 9.2M PN entities have been detected with LIA_NE corresponding to 811K different forms, further reduced to 220K by removing PN occurring less than 4 times
- a clustering method has been applied to this 220K list by associating forms sharing common substrings and occurring in the same newswire
- finally a set of 110K clusters has been obtained, each of them representing several forms of a person name; the normalized form chosen is the most frequent sequence *firstname+lastname* in the cluster.

3. Error recovery approach

Transcription errors on Person Names, be them caused by Out of Vocabulary (OOV) words, by unseen context or by any possible cause, are likely to result in transcribed words that are phonetically close to the original name. It can be a simple substitution (*Karim examen* instead of *Karim Benzema*, *Marc librement* instead of *Marc Lièvremont*) or a sequence of erroneous short words (*ou ma Thurman* instead of *Uma Thurman* or *Calais les valses* instead of *Cadel Evans*). In this paper, we propose to recover from such errors by comparing the phonetic representation of recognized words to a PN dictionary.

The error recovery strategy consists in: first *detecting* Person Names error regions and then *searching* for Person Names in error regions on the basis of their phonetic representation.

3.1. Error detection and characterization

In previous work [1], we have proposed an approach for error region detection and characterization in LVCSR transcriptions. It is a well-known phenomenon that LVCSR errors tend to appear in consecutive error regions. Beyond the classical confidence measure approach that estimates the probability of each word to be correct, we consider error detection as a segmentation task. Furthermore, simply detecting error regions is not enough for general purpose and it is important to characterize them. We have proposed to characterize error region with respect to the nature of the words that yielded the error region. Characterization is done along four error classes: *Person Names, Other Proper Nouns, Homophones, Others.* Driven by applicative considerations, we believe that this approach is more suitable to define relevant strategies in order whether to ignore errors (e.g. *Homophones*) or to try to correct them.

In this paper we combine two of the approaches described in [1] and we focus on the detection of Person Names error regions.

Sequential approach

The segmentation and characterization steps are applied sequentially. Error detection is simply performed thanks to a threshold on the ASR confidence measures. Segmentation is performed by gathering consecutive errors to form an error region. Then a supervised classifier (`icsiboost`[1]) is applied on error regions with various features representing the region itself (word bigrams, POS trigram and syntactic chunks[2], number of words, average number of syllables per words) and its context (quadrigram on the 5 previous words, duration and average confidence measure of the speaker turn).

Integrated approach

Error regions have specific properties depending on their nature (various average length for instance) suggesting to use integrated approaches in order to simultaneously segment and categorize error regions. This is achieved thanks to Conditional Random Fields (CRF) with an IUO underlying model (Inside, Unique, Outside) and the following feature set: word bigrams POS tags and syntactic chunks, confidence measure and duration of current, previous and next word.

3.2. Error correction

Given an automatically detected error region, the error correction process consists in searching in a Person Name dictionary for the entity whose phonetic representation is the most similar to the phonetic sequence associated to the

[1] http://code.google.com/p/icsiboost

[2] http://pageperso.lif.univ-mrs.fr/~frederic.bechet/download.html

recognized words. This approach was originally studied in the context of open-vocabulary spoken term detection [9]. In the general case, such an approach can be efficient in terms of spoken term detection recall but has to be carefully used in order to avoid low precision rates. In this paper, we show that it can be a relevant approach for searching PN provided that the search step is guided by a preliminary error detection step.

Constitution of the PN database

In order to be consistent with the further Person Named Entity detection and Linking stages, the dictionary is extracted from the one described in section 2.3 and is composed of the 20K most frequent PN. Each entry is associated to its phonetic representation thanks to the Orange Labs grapheme to phoneme converter (G2P). We keep pronunciation variants for each entry, leading to a total amount of 135K phonetic sequences dictionary.

Phonetic search

Given a detected error region $E=\{w_1,\ldots,w_n\}$ of n consecutive words, correction is achieved along the following process:

- The sequence of n words is transformed into a sequence of phonemes, preserving the inter-word coarticulation phenomena (which are particularly usual in French with the *liaison* phenomenon). The G2P providing phonetic variants, we keep the longest phoneme sequence as our phonetic search space $\Phi=\{\varphi_1,\ldots,\varphi_k\}$.
- Phonetic search is performed thanks to dynamic programming through the phoneme sequence Φ. It looks for optimal alignment (i.e. with minimum distance) between the sequences of phonemes of the PN database entries and any sub-sequences within Φ. The alignment distance is defined as the sum of the costs of the operations (phoneme pairs substitution, deletion and insertion) involved in the alignment. It is then normalized by the number of operations. The insertion, deletion and substitution costs are derived from a pre-computed phoneme confusion matrix.
- The result of the phonetic search step is an *n-best* list of potential Person Names matching with a sub-sequence of phonemes from Φ and ranked according to their normalized distance to the matched sub-sequence.
- The previous matching scores do not reflect length constraints. It is then possible that a very short name matches perfectly a subsequence of one or two phonemes from the original Φ sequence. In order to alleviate this problem, a re-ranking process is applied to include coverage constraints. It allows favoring the detection of a longer name with a higher distance provided that this distance ranges in a fixed delta from the one-best distance.
- Finally, it can happen that several Person Names are mentioned in a single PN error region. The previous step is iterated as long as non-overlapping hypotheses can be produced from the *n-best* list.

Implementation

Using a phoneme confusion matrix allows taking into account the fact that some phonemes are more likely to be omitted or inserted than others and that some phoneme pairs are more likely to be confused than others. It is computed maximising the likelihood of the alignment between a reference phoneme sequence obtained from manual transcriptions and the one obtained from LVCSR output on the same training corpus. This learning step is performed iteratively by the EM algorithm.

In practise, it can happen that only the last name yields an error region while the first name is correctly transcribed (eg. a common first name and an unknown last name). In order to increase the precision of the search process, we systematically append the error region with the previously transcribed word when this word is tagged as a first name by the POS.

The next section shows how this error correction approach, coupled with a suitable error detection process can improve the overall PNR process.

4. Experiments

4.1. Database description

The error region detection and characterization models have been trained on a 14 hours TV Broadcast News shows corpus composed of 38 shows from 7 French generalist channels [8]. The phoneme confusion matrix, for our 33 phonemes set, has been trained on the same corpus: 521k reference phonemes were aligned with 555k hypothesized phonemes.

PNR experiments are run on a corpus of TV shows from two French channels (2 shows from BFMTV and 5 shows from LCP, none of them are part of the training corpus) provided by the French REPERE evaluation program. The corpus is characterized by a strong variety in terms of topics and types of shows (debates, extracts of parliament allocutions, reports, news…). A development corpus (DEV) is used to tune the error recovery strategy and a test corpus (TEST) is used to assess to overall PNR performance. DEV and TEST both consist of 3 hours of speech from about 30 extracts of shows. PN entities have been manually annotated and associated with a normalized linked form. Our entity linking approach can associate a last name to a reference form but does not perform contextual reference resolution (isolated first name are not linked to a reference form). As a consequence, we discard from the database persons that are only mentioned by their first name.

	DEV	TEST
#words	34,312	34,683
word error rate	20.8%	24.4%
#error regions	2,296	3,069
average length	2.9	2.8
#PN	581	430
# PN error regions	234	184
average length of PN error regions	4.2	3.8

Table 1 *REPERE corpora description*

The proportion of PN occurrences covered by the dictionary is 97.2% for DEV and 95.8% for TEST. 13.9% of PN in DEV contain at least one Out-of-Vocabulary (OOV) word with respect to the ASR lexicon (16.7% for TEST). OOVs account for 38.1% of PN error regions for the DEV corpus (43.3% for TEST) confirming the interest of characterising PN error regions in general and not only from the OOV point of view.

4.2. Impact of error recovery strategies

In order to quantify the impact of the error recovery strategy on the overall PNR task, we measure *PNR recall* and *PNR precision*. *PNR recall* is the number of normalized Person Names that are correctly recognized, over the total amount of reference normalized PN. *PNR precision* is the equivalent value

over the total amount of detected normalized PN. In Figure 1, *PNR recall* is plotted against *PNR precision* by varying the threshold on the tolerated distance between the original phoneme sequence and the hypothesized names phoneme sequences.

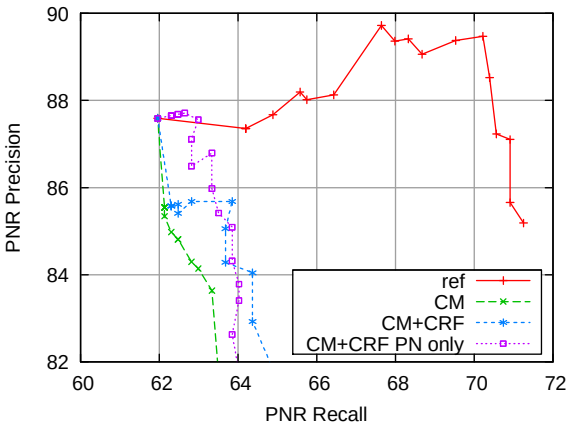


Figure 1: *Person Name Recognition Recall and Precision for various error region detection strategies.*

The top curve (*ref*) corresponds to the optimal case where PN error regions are manually segmented. It illustrates the interesting potential of the proposed correction method and provides *Oracle* PNR performance with the whole process including our correction paradigm. Starting with a 62.0% *PNR recall* and a 87.6% *PNR precision*, the curve shows that it is possible to reach an optimal 70.2% *PNR recall* with an improved 89.4% *PNR precision*.

Systematically searching PN in automatic transcriptions would necessarily yield poor performance. Contrastive experiments were achieved for the sake of comparison and leaded to around 30% absolute drop in *PNR precision* even with a very strict tolerance on the distance confirming the need for an error detection strategy in order to guide the correction process. The three other curves presented in Figure 1 correspond to three different error region detection strategies. As a baseline, we directly used the word level Confidence Measure (CM) provided by the ASR engine. If several consecutive words have a CM below a given threshold (see [8] for threshold optimisation) they are considered as an error region. Without any further selection process, this strategy (illustrated by the *CM* curve) eventually allows increasing the *PNR recall* but with a significant drop in *PNR precision*.

Additionally using a CRF to detect error regions ("*CM+CRF*" through a simple OR fusion) improves the quality of the overall process. In fact, it can happen that a misrecognized PN generates a succession of several erroneous words among which one can be a short word with a high CM. Simply applying a threshold results in multiple error regions, thus preventing the correction process to be applied on the relevant phonetic support. Due to the rich modelling provided by the structural and contextual features, CRFs are more likely to detect a single error region, even if all recognized words do not have a CM below the threshold. As a result, *PNR recall* is higher for the *CM+CRF* curve. However, the *PNR precision* is still degraded.

The last experiment illustrates the importance of the error region characterization step. Following the method described in section 3.1, we only apply the correction strategy to those error regions that are automatically characterized as error regions generated by misrecognized person names. The "*CM+CRF PN only*" curve shows that the *PNR recall* can be increased to a certain extent without any loss in *PNR precision* (following the *Oracle* curve).

In order to validate the approach on the TEST corpus, operating points have been chosen in order to optimise the F-measure for the "*ref*" and "*CM+CRF PN only*" curves.

TEST	No correction	Manual error reg.	Automatic error reg.
PNR Recall	62.1	68.6	63.3
PNR Precision	80.7	82.6	81.0
overall w.e.r.	24.4 %	24.1 %	24.3 %

Table 2: *Performances on the TEST corpus*

The second column in Table 2 confirms the potential of the error correction approach when applied on manually segmented error regions and the third column confirms the performance of the fully automatic approach with an increasing recall and precision for equivalent overall word error rate (w.e.r.). When focusing on the speaker turns where a correction has been applied, the w.e.r. decreases from 23.1% to 22.1% for the manual error segmentation approach and it decreases from 18.3% to 18.0% for the fully automatic approach.

5. Conclusion

We have proposed an error recovery approach dedicated to errors on Person Names (PN). It has been integrated in a complete Person Name Recognition (PNR) task (including ASR, Named Entity Recognition and Person Entity Linking) and evaluated on a TV shows corpus available from the REPERE evaluation program. The error correction approach, consisting of searching for PN within the phonetic representation of wrongly recognised words, proved to be very promising when applied to the true (manually detected) PN error regions. When coupled with an automatic PN error region detection system, we were able to increase the overall PNR recall without degrading PNR precision nor the overall word error rate. Further improvements in the difficult error region detection and characterization task should yield even better results.

6. References

[1] Lecorvé, G., Gravier, G. and Sébillot, P. "An unsupervised Web-based topic language model adaptation method", ICASSP'08, Las Vegas, 2008.

[2] Rastrow, A., Sethy, A. and Ramabhadran, B., "A new method for OOV detection using hybrid word/fragment system", ICASSP'09, Taipei, 2009.

[3] Béchet, F., Favre, B. and Damnati, G. "Detecting person presence in TV shows with linguistic and structural features", ICASSP'12, Kyoto, 2012.

[4] Jousse, V., Petit-Renaud, S., Meignier., S., Esteve, Y. and Jacquin, C., "Automatic named identification of speakers using diarization and ASR systems", ICASSP'09, Taipei, 2009.

[5] Gauvain, J.L., Lamel, L. and Adda, G., "The LIMSI Broadcast News Transcription System". *Speech Communication*, 37(1-2):89-108, 2002.

[6] Béchet, F. and Charton, E., "Unsupervised knowledge acquisition for extracting named entities from speech", ICASSP'10, Dallas, 2010.

[7] Stern, R., Sagot, B. and Béchet, F., "A Joint Named Entity Recognition and Entity Linking System", Proc. Workshop on Innovative hybrid approaches to the processing of textual data, EACL 2012, Avignon, 2012.

[8] Dufour, R., Damnati, G. and Charlet, D. "Automatic error region detection and characterization in LVCSR transcriptions of TV news shows", ICASSP'12, Kyoto, 2012.

[9] Dubois, C. and Charlet D., "Using textual information from LVCSR transcripts for phonetic-based spoken term detection", ICASSP'08, Las Vegas, 2008.