Person Name Recognition and Linking from Overlay Text in TV Broadcast Shows

Geraldine Damnati\textsuperscript{1}, Benoit Favre\textsuperscript{2}, Frederic Bechet\textsuperscript{2}, Delphine Charlet\textsuperscript{1}

\textsuperscript{1}Orange Labs, Lannion, France \{geraldine.damnati,delphine.charlet\}@orange.com
\textsuperscript{2}Aix Marseille University LIF-CNRS \{benoit.favre,frederic.bechet\}@lif.univ-mrs.fr

Abstract

Identifying people in video broadcast is by nature a multimodal task: persons can be identified thanks to biometric information (face or voice), or thanks to a reference to their identity in the overlaid text or the speech content. In the framework of the French evaluation program Repere, this paper presents a method for identifying speakers in videos without any a-priori models, based only on overlaid text often used to introduce guests or journalists occurring for the first time in a given TV show. We show that Entity Linking improves speaker identification performance by reducing ambiguities in OCR transcriptions and allowing to add biometric constraints in the multimodal fusion process. All the methods presented are evaluated on the Repere video corpus of broadcast shows from 2 French TV channels and 5 different shows (news, talk shows, magazine).

Index Terms: OCR, Named Entity, Entity Linking, Multimodal fusion

1. Introduction

Identifying people in video broadcast is by nature a multi-modal task: persons can be identified thanks to biometric information (face or voice), or thanks to a reference to their identity in the overlaid text or the speech content. In addition to finding the name of persons appearing in videos, it is also important to link them to identities from a database, ensuring that the identity is not ambiguous, and potentially leveraging linked metadata. From an applicative point of view this tasks is interesting as it allows multi-modal queries in video archives, such as “find all the shows where X was a guest”. From a scientific point of view, this task is an opportunity to study multi-modal fusion algorithms as well as methods for improving information extraction from speech and images thanks to combined decoding in both modalities.

The Repere challenge consists in identifying persons in video shows using cues from spoken content (speaker identity and words), and video content (faces and overlaid text) [1]. Systems participating in the challenge must generate a list of segments with person names according to the presence of persons mentioned in the visual and audio modalities, using both biometric models and context analysis. The challenge provides a set of videos manually annotated with speaker segmentation, speech transcription, overlaid text transcription and face outline. All image-related annotations are sampled every 10 seconds on so-called key-frames.

In the framework of the French evaluation program Repere, this paper presents a method for identifying speakers in videos without any a-priori models, based only on overlaid text often used to introduce guests or journalists occurring in a given TV show. This method consists of four steps:

1. Overlaid text detection and Optical Character Recognition (OCR).
2. Person name detection in the OCR transcriptions (Named Entity tagging).
3. Identity finding from the person name detected (Entity Linking).
4. Speaker Identification (multimodal fusion with the output of a speaker diarization module).

We show in this paper that Entity Linking improves speaker identification performance by reducing ambiguities in OCR transcriptions. Entity Linking can also be used to add additional constraints to the multimodal fusion process such as gender and age, although this was not implemented in the experiments reported in this paper.

This paper is structured as follows: Section 2 presents some related work on OCR in videos as well as entity detection and linking; Section 4 describes the OCR module used in this study, based on a multi-frame decoding and fusion algorithm; Section 3 presents the construction of knowledge-based resources used for linking names to database entries; Naming speakers from the knowledge base is described in Section 5; Finally, the complete process is evaluated on the Repere video corpus, containing 48 hours of broadcast shows from 2 French TV channels and 5 different shows (news, talk shows, magazine), in Section 6.

2. Related work

Optical Character Recognition (OCR) is a well studied topic for which very effective solutions have been created (see [2] for a review). In the Natural Language Processing community, OCR output is considered as a noisy source of text in the same respect as Automatic Speech Recognition (ASR) transcripts, and it has been used in many evaluation campaigns, on various NLP tasks, to test the robustness of involved systems. In particular, the Automatic Content Extraction (ACE) campaigns tackled named entity mention and relation extraction on OCR output [3]. Unlike systems designed for the ACE task which have a hard time locating entity mentions [4], person name locations are relatively easy to find in overlaid text, but linking names to database entries, our focus, requires very good character accuracy in the detected names. Practitioners have also studied the impact of OCR and ASR errors on the Named Entity Recognition (NER) task and how to limit performance degradation due to errors [5]. In [6] it has been shown that NER performance on OCR output decreases by about half a point of F-score for each additional point of word error rate.
In order to limit those errors, multiple hypotheses can be retained for each character, forming a lattice which is processed by the NER system [7]. In videos, overlaid text appearing on a changing background might generate different errors on consecutive frames, which is exploited by generating a temporal confusion network with more accurate oracle solution [8]. The problem of information extraction from erroneous OCR of old newspaper prints was addressed by [9]. In addition to using a dictionary, they use hand-crafted letter sequence substitution to correct out-of-vocabulary words. [10] compare ASR and OCR for retrieving videos in the framework of the TRECVID evaluations, and find OCR to be more reliable than ASR.

Here, our experimental validation is motivated by the task of finding speaker names, which has mainly been tackled by looking for direct cues from ASR transcripts [11, 12]. For instance, [13] use decision trees to decide whether an uttered person name belongs to the current speaker, the previous speaker, the next speaker or a third party. Such approach has been applied on radio broadcasts, however on TV content there are other sources of information than just ASR that can be used to find the task, such as overlaid texts.

Naming persons from OCR is a particular application of information extraction from videos which has been tackled in a few projects. The Name-it system [14] extracts names from OCR transcripts of videos and associates them with faces which cooccur the most with them. [15] associate names with TV-series characters by relying on the script (where speech is attributed to speakers) and lip activity detection. More recently, [16] describe a multimodal system for naming people in TV broadcasts which uses OCR output to retrieve names from overlay texts. This system relies on an unsupervised method for naming speakers described in more details in [17] which is the most comparable to our work. We tackle the problem in the same experimental framework but we extend the reach of our research by linking speakers to a person database and study its benefits on naming speakers.

In the next sections, we present details of the proposed method for finding names in OCR, linking them to a knowledge base and using the results to name speakers.

### 3. Knowledge base extraction from semi-structured data

In order to be able to link a name with an identifier referring to a unique person, two resources are needed: firstly a dictionary of person entities with a description for each entry identifying a person without any ambiguity (date and place of birth, sex, citizenship, occupation, etc.); and secondly a list of acceptable surface forms for referring to each person (first name+last name, title+last name, nickname, etc.).

The first resource can be found in databases made available on the WEB, such as DBpedia. In these databases the coverage is usually quite good for historical figures, and famous people linked to the current news in domains such as politics, art, sport or science. However these resources do not contain regular people appearing sporadically in broadcast content like witnesses and actors of an unexpected events or experts on a given topic. Therefore these semantic web databases have to be extended with additional person names coming from news data of a given time period and a given broadcast channel. For the second resource, alternate surface forms for a given person name can be either generated or collected. Generating alternate forms can be done thanks to a set of rules, however it will lead to an over generation of forms and it prevents obtaining nicknames that can’t be produced automatically. In this study we used a single process to collect new entities related to the news of a given time period and alternate realizations for each of them.

This process can be described as follows:

- A text corpus of newswire corresponding to the dates of the broadcast shows to process is first collected.
- A Named Entity tagger (LIA,NE [18]) is then applied to this corpus in order to detect mentions of person names in each document.
- All person name mentions occurring in the same newswire and sharing at least one word are grouped together in the same cluster.
- Once all the corpus has been processed we merge the clusters sharing at least one identical mention.
- Finally the clusters are filtered by using a set of heuristic rules and removing low frequency items.

At the end of this process we obtain a set of clusters, each of them representing a single person with a list of associated name mentions and frequency. For example, for referring to Carla Bruni-Sarkozy, we obtain the cluster presented in table 1.

We applied this process to a large corpus of newswire spreading from 2004 to 2012. About 9.2M NEs were automatically detected. After the clustering and filtering process we obtained 117K clusters containing 162K mentions of person names. The last step of this process is to link each cluster with a unique identifier. We used for this purpose the entity database ALEDA [19], which is a structured version of Wikipedia. ALEDA contains about 225K person entities. For each of them we have the gender as well as a small text description extracted from the Wikipedia web page. The linking process is simply done by checking for each cluster if there is a match in ALEDA for one of the name belonging to this cluster. If there are several matches, we keep the match with the highest number of occurrences in the newswire corpus. From the text description of each person found, we extract 6 characteristics (when available): gender, date of birth, date of death, citizenship, main topic attached to the person (media, sport, ...) and a boolean indicating if the person can be expected to speak French or not. These characteristics are automatically extracted thanks to heuristic rules defined on ALEDA. If one or several characteristics are not available, they remain undecided.

### 4. Overlay Person Name recognition

The Overlay Person Name (OPN) recognition process is made of 3 steps in our approach: text box detection; Optical Character Recognition producing a confusion network of characters; person name recognition in the character hypotheses.

<table>
<thead>
<tr>
<th>Name</th>
<th>NEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carla Bruni-Sarkozy</td>
<td>1180</td>
</tr>
<tr>
<td>Carla Bruni</td>
<td>704</td>
</tr>
<tr>
<td>Mme Bruni-Sarkozy</td>
<td>141</td>
</tr>
<tr>
<td>Carla Sarkozy</td>
<td>67</td>
</tr>
<tr>
<td>Carla Bruni Sarkozy</td>
<td>16</td>
</tr>
<tr>
<td>Mme Sarkozy</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 1: Example of cluster obtain with the alternate name mentions for Carla Bruni-Sarkozy
Text box detection is achieved with a convolutional neural net approach described in [20], applied to raw pixels without any chromatic pre-processing. It is run at each frame and text boxes are searched in a predefined region of the image corresponding to the area where Person Names are likely to be overlaid (usually at the bottom of the image).

The OCR process performed on the detected boxes uses Tesseract\(^1\), a standard open-source OCR system. In order to account for inaccuracy of text detection boxes, we configure Tesseract to explore small variations (±2 pixels) of the boxes and retain the one for which it is most confident in the sequence of predicted characters. Spurious detections are avoided by filtering out transcripts that contain more non-alpha-numeric characters than regular characters.

Text is tracked on consecutive frames by comparing the position and size of detections. Then sequences of characters recognized at each frame along a given track are merged in a Confusion Network (CN) in order to make the best of potentially different transcripts due to background variations. The process incrementally creates a hypothesis graph by aligning observed character sequences using Levenstein alignment to a pivot sequence. For pivot, we choose the most common character sequence in the track and then successively align other sequences by decreasing frequency. To prevent a large amount of insertions and deletions, the alignment cost matrix is tuned to the following costs: insertions and deletions have a cost of 150, substitutions have a cost of 100. In addition, highly confused characters pairs such as i ↔ I, v ↔ V, s ↔ S, p ↔ P... are given a cost of 1. This results in much more accurate confusion networks.

The posterior probability of a character at a given position of the CN corresponds to the frequency of this character at this position relatively to the total number of frames of the track. Epsilon transitions are added in order for posterior probabilities to sum up to one at each positions. Normalizing by the total number of frames guarantees that positions with only a few inserted characters will not be kept in the best sequence hypothesis.

Extracting person name from the best hypothesis of the CNs of the OCR system is not straightforward. Firstly the selected potential text area can contain several types of informations apart from the name of the speaker. It can be the title of a report, the subject or the location of the report. In order to discriminate predictions apart from the name of the speaker, it can be the title of the subject or the location of the report. In order to discriminate OPNs from other overlaid information, a rule-based detection module has been designed. A text track is considered as an OPN if it is above another text track, if the height of characters is bigger than the height of characters in the text track underneath, and if it contains less than four words. Secondly, once a name is detected, some characters might have been misrecognized, leading to a failure in the linking process. To overcome these OCR errors we use the person name lists presented in section 3 as a language model to correct the OCR hypotheses. These lists are turned into a Finite State Transducers (FST) taking as input the characters contained in the names, and outputting as result the name IDs. The output of the OCR process is also turned into an FST, the multiple paths corresponding to alternative character strings by adding/removing or replacing characters output by the OCR. Each distortion has a cost and by composing this FST with the person name list FST and calculating the best paths, we obtain the best match between all the different person names of our resources and the output of the OCR. To each match is associated a distortion score, corresponding to the Levenshtein distance between the OCR output and the person names. When no match is found, or if the distortion score is too high, the best path in the OCR confusion network is kept. The decision process presented in the next section is in charge of further filtering these name hypotheses.

5. Decision process and speaker identification

The decision process consists in validating or rejecting the OPN hypotheses output by the OCR process. For each text box, one person name hypothesis $H$ is output. This hypothesis can be either the best match found in the person name FST or the best character string in the OCR confusion network if no match is found. The decision process is based on the following linking process:

- If $H$ has a very high confidence and belongs to one of the person name clusters of the database presented in section 3, we link $H$ with the canonical form of the name in the cluster;

- if $H$ matches a name in the person name FST but with a low confidence and if an internet search of the exact original string yields a number of hits higher than a threshold (arbitrarily set to 400 in our experiments), we keep the original form without linking it to the database;

- In all the other cases the name is likely to be spurious and is therefore discarded.

The speaker identification process consists of the joint processing of speaker diarization output and OPN hypotheses. In order to evaluate our entity detection and linking method, not only on the capacity to retrieve a given name, but also on the global task of identifying people in videos, we develop a very simple speaker identification process: the assumption is made that if a name occurs in overlaid text when a person speaks, it is the name of the speaker. In our evaluation corpus, this assumption is true in 80.4% of the cases. Based on this assumption, a speaker identification system relying on overlaid text is proposed. First, speaker diarization is performed. It consists in detecting speech from non-speech, segmenting speech into turns (each turn containing the speech of a single speaker), and grouping together speech turns from the same speaker into speaker clusters. This diarization step is performed with the system presented in [21]. Then, the speaker identification system tries to give a name to each speech turn, as follows:

For turn $t$, let $opn\_name(t)$ be the OPN which has maximal overlap duration with $t$, and $opn\_name(Cluster(t))$ the OPN which has maximal duration with the speaker cluster containing $t$. Then, each turn is named with $opn\_name(t)$ if defined, or else with $opn\_name(Cluster(t))$ if defined, or otherwise it remains unnamed. Both name sources can be undefined when no OPN was detected during the corresponding speech.

6. Experiments

All the experiments reported in this paper have been performed in the framework of the Repere challenge. We consider here the unsupervised track of this challenge, which precludes the use of biometric models (voice or face), in order to focus on person identification thanks to OPNs.

In our experiments, we use the Repere phase1 corpus in a leave-one-out fashion. It consists of 135 videos totaling over 48 hours of air time in which about 24 hours have been annotated with speaker segmentation (including overlaps) and speaker

\(^1\)https://code.google.com/p/tesseract-ocr/
names. Even though all data of the audio modality is annotated, the visual modality is only annotated through a set of keyframes chosen roughly every 10 seconds (8624 keyframes for the phase 1 corpus). Therefore, although our OPN detection and speaker identification systems are applied to the whole 48 hours corpus, they can only be evaluated on the keyframes for which the ground truth is available. On these keyframes, 80.4% of the OPN correspond to the current speaker.

The OPN recognition results are presented in table 3 according to 3 conditions:

- **OCR**: the OPN hypotheses are the best character strings obtained on the OCR confusion networks;
- **OCR+FST**: in this condition the person name Finite State Transducer with the \(L_4\) list is composed with the OCR character strings, as presented in Section 4;
- **OCR+FST+linking**: the decision process described in Section 5 is applied to the output of the OCR process. We discard all OCR hypotheses that can’t be linked, either in the person database or thanks to internet queries.

A we can see in table 3, using a large person name list in order to correct OCR output brings a large improvement in the OPN recognition as well as the speaker identification performance. The linking process has an impact mostly for the speaker identification task. This means that increasing the OPN precision is crucial in order to prevent propagating wrong identities that can spread on large portions of audio signal.

Table 2: Coverage of the person lists on the Repere corpus

<table>
<thead>
<tr>
<th>Role</th>
<th>(L_1)</th>
<th>(L_2)</th>
<th>(L_3)</th>
<th>(L_4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>42.5</td>
<td>77.5</td>
<td>88.1</td>
<td>89.9</td>
</tr>
<tr>
<td>R1</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>R2</td>
<td>90</td>
<td>90</td>
<td>90</td>
<td>90</td>
</tr>
<tr>
<td>R3</td>
<td>69.2</td>
<td>69.2</td>
<td>69.2</td>
<td>69.2</td>
</tr>
<tr>
<td>R4</td>
<td>15.5</td>
<td>63.8</td>
<td>81.0</td>
<td>82.8</td>
</tr>
<tr>
<td>R5</td>
<td>25.9</td>
<td>77.0</td>
<td>88.1</td>
<td>91.1</td>
</tr>
<tr>
<td>Size</td>
<td>96</td>
<td>7K</td>
<td>163K</td>
<td>425K</td>
</tr>
</tbody>
</table>

Table 3: Overlay Person Recognition and Speaker Identification results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>OCR</td>
<td>63.2</td>
<td>67.0</td>
<td>65.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ FST</td>
<td>90.6</td>
<td>70.2</td>
<td>79.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ linking</td>
<td>91.2</td>
<td>70.7</td>
<td>79.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spk ident.</td>
<td>71.3</td>
<td>59.9</td>
<td>65.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OCR</td>
<td>71.3</td>
<td>59.9</td>
<td>65.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ FST</td>
<td>77.9</td>
<td>65.4</td>
<td>71.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ linking</td>
<td>84.3</td>
<td>67.6</td>
<td>75.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4: % of correct speaker identification according to speaker Roles with 2 sizes of person lists for the linking process

Future work includes finer use of meta-data to constrain speaker identification and better analysis of the context in which a superimposed name should be propagated to the current speaker.

7. Conclusion

This paper tackles the problem of unsupervised speaker name attribution from overlaid text in TV broadcasts. We present a system which uses optical character recognition to retrieve potential person names, link them to a knowledge base, and use meta-data from this database to constrain speaker identification. This entity linking procedure leads to an improvement in identification performance compared to blindly propagating OCR output, as measured on videos of the Repere corpus.

8. Acknowledgements

This work is partially funded by ANR under project PERCOL 2010-CORD-102-01.
9. References


