www.axes-project.eu

Project acronym	**AXES**
Project full title	**Access to Audiovisual Archives**
Project No	**269980**
	Large-scale integrating project (IP)

Deliverable D5.5

Update on final software toolbox for extracting spoken and written entities

September 2014

SEVENTH FRAMEWORK PROGRAMME

Objective ICT- 2009.4.1: Digital Libraries and Digital Preservation

PROJECT DELIVERABLE REPORT

Project	
Grant Agreement number	*269980*
Project acronym:	*AXES*
Project title:	*Access to Audiovisual Archives*
Funding Scheme:	*Large-Scale Integrating project (IP)*
Date of latest version of Annex I against which the assessment will be made:	30.6.2014

Document	
Deliverable number:	D5.5
Deliverable title	Update on final software toolbox for extracting spoken and written entities
Contractual Date of Delivery:	30/09/2014
Actual Date of Delivery:	30/09/2014
Author (s):	Christoph Schmidt (FHG)
Reviewer (s):	Yann Mombrun (Airbus DS)
Work package no.:	5
Work package title:	Spoken and Written entities
Work package leader:	FHG
Version/Revision:	V0.2
Draft/Final:	Final
Total number of pages (including cover):	18

CHANGE LOG

Reason for change	Issue	Revision	Date
Initial version (FhG)	1	0	19/09/2014
Minor update (Airbus DS)	2	0	22/09/2014

DISCLAIMER

This document contains description of the AXES project work and findings.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any responsibility for actions that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of the AXES project and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 27 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors. (http://europa.eu.int/)

AXES is a project partly funded by the European Union.

TABLE OF CONTENTS

SUMMARY

This document describes the different modules that are included in the updated WP5 final toolbox and that will be integrated either in the web services of the V2 version of the AXES system. The descriptions in this document build upon the previous deliverables D5.4 (Report on final software toolbox for extracting spoken and written entities).

INTRODUCTION

The goal of WP5 "Spoken and Written Entities" inside the AXES project is the reliable detection of entities from audio and video streams. The various modules used in this process may use previously available metadata in the form of associated textual data. This textual data comprises production metadata like scripts, lists of actors or directors, or summaries, as well as user generated information like comments or tags.

The main outcome of WP5 processing a multimedia asset is an enriched metadata object following the format defined in the AXES consortium. It contains entity annotations, where entities can be persons, locations or buildings, or events. These annotations are either derived directly from the audio through speech recognition and subsequent entity extraction, or from the video through video OCR and entity extraction.

In the end these annotations are then used by WP6 for managing links between documents as well as building a search index.

MODULE ANALYSIS

M5.1 Speech Recognition for German (FHG)

Objective

This module is responsible for performing automatic speech recognition (ASR) and thus creating textual transcriptions of German audio content. These transcriptions are then processed by the entity extraction component and finally indexed for direct retrieval.

Description

The systems for German and English ASR are based on the Kaldi-toolkit [1] which uses HMMs with Deep Neural Networks (DNNs). DNN-HMMs evaluate the fit between the frames of features representing the acoustic input and the HMM states using a feed-forward neural network that takes several frames of features as input and produces posterior probabilities of HMM states as output. DNNs use multiple hidden layers (that is why they are called "deep") of non-linear hidden units and a very large output layer to accommodate the large number of HMM states, because each phone is modelled by a number of context-dependent triphone HMMs.

Before running the ASR system, a pipeline of components performing audio segmentation and speech detection is executed. The language model is based on newswire reports and various other web-based text sources and trained using the IRSTLM language modeling toolkit [2].

For German, FHG used 636 hours of manually annotated acoustic data from the GER-TV1000h corpus [3]. Moreover, for this setup a recurrent neural network language model rescoring [4] was applied. This led to a reduction in the word error rate (WER) on the DiSCo evaluation benchmark set [5] from 22.4% (last year's system) to 16.3% using DNNs and the increased training data size and further to 15.2% by applying RNN rescoring.

Integration status

This component has been integrated in the AXES system.

AXES OPEN status

This component will not be part of AXES OPEN.

M5.2 Speech Recognition for English (FHG)

Objective

Comparable to M5.1, this module creates a textual representation of English audio content.

Description

For the English Automatic Speech Recognition (ASR) system, FHG uses the same training and decoding system as for German ASR. For English the models are built based on the following training material:

For the acoustic model, the following corpora are used:

- The 1996 Broadcast News Speech Corpus (HUB-4 '96, as described in LDC97T22 and LDC98S71), which contains a total of 104 hours of broadcasts from ABC, CNN and CSPAN television networks and NPR and PRI radio networks with corresponding transcripts, and

- The Boston University Radio Speech Corpus (BURSC, as described in LDC96S36), which consists of professionally read radio news data, including speech and accompanying annotations.

For the language model, we use the transcripts of BURSC and HUB-4 '96 and the TIPSTER data collection (LDC93T3A), which consists of 5.2M phrases. For the dictionary, we use the CMU Pronouncing Dictionary (CMUDict) version 0.7a as released in February 2008. For unknown words, we use Sequitur, which is a GPL phonetizer as described in [6]. It does not come with a predefined model; rather, the phonetic transcription has to be learned from existing models (based on CMUDict).

Using the new DNN-based decoder, FHG further improved the word error rate from 22.9% to 20.5%.

Integration status

This component has been integrated in AXES system.

AXES OPEN status

This component will not be part of AXES OPEN. Instead, the open-source decoder Sphinx (http://cmusphinx.sourceforge.net) will be integrated as an alternative.

M5.3 Speech Recognition for Dutch (UT)

Objective

In contrast to the other ASR modules, Dutch audiovisual content is analyzed outside of the processing chain. The results are then injected into the system in the form of subtitles, providing a transparent solution in the case when no manual subtitles exist for a given recording.

Description

Airbus DS has provided a WebLab service that enables to find the srt file (generated by UT) associated to a given document, to read it and to add the text segments to the document.

Integration status

The component in charge of reading srt files has been integrated in AXES system.

AXES OPEN status

The Dutch ASR is not part of AXES OPEN; but the component in charge of reading SRT file will be integrated in the OPEN AXES package.

M5.4 Forced Alignment on Word Level in Dutch (UT)

Objective

The forced alignment (FA) module synchronizes textual representations of speech, usually referred to as 'speech transcript', with speech in audio data.

Description

Since usually manual speech transcriptions differ noticeably from the actual spoken words, special care has to be used when aligning these to the audio stream. Since we do not foresee to process large amounts of manually transcribed material, the availability of this component will not affect the analysis results to a great extent.

AXES OPEN status

The component will not be part of AXES OPEN.

M5.5 Video OCR (KUL)

Objective

Text that is visible in video recordings, either in the form of inserts or as actual filmed text, is recognized by this module and converted into machine-readable annotations.

Description

Two different approaches to video OCR have been investigated: one based on text edges developed by partner KUL and one based on character detection, developed by partner INRIA.

The text edge-based method developed by KUL earlier turned out to be too computationally expensive for use on a large scale. Therefore, it was decided to change the pipeline, now starting also from characters as detected by a (modified) MSER. This allows to quickly reduce the search space and to spend most processing time on regions that are likely to contain text. Pairs of characters are then further analysed and classified as text or non-text based on gradient statistics, stroke width statistics, as well as relative position of the characters. Characters are grouped into text lines. We rely on Generalized Extreme Value Distribution to determine the optimal threshold for binarizing the image based on difference of gamma functions.

Character-based word detection and video OCR

In parallel, partner INRIA developed a scheme for video OCR starting from characters, as detected with the MSER interest point detector [7]. This scheme runs faster, and also has the actual OCR included. Therefore, this scheme was used for the TrecVid MED evaluations. The INRIA text detector in videos is based on the assumption that each text character can be detected by an extremal region, as for example in [8]. The Maximally Stable Extremal Region (MSER) detector [7] is a fast and reliable way of extracting such regions. Here we use the luminance channel to extract MSERs, as the saturation and hue channels typically undergo much stronger encoding distortions in videos than in still images. Given a set of extracted MSER the goal is to connect them into words. Similarly to [8], we use the following pipeline: First, we remove MSERs that do not have a suitable aspect ratio, are too small or too large, and have weak gradients on their boundary. They are, then, grouped into pairs according to their relative position. Pairs of regions with too big of a difference in color or stroke width are eliminated. Regions that passed the previous steps are then grouped into text lines. Furthermore, we train a classifier to split the lines into distinct words, as detecting inter-word spaces is difficult due to the variability in existing fonts and the presence of noise in the region masks. Finally, MSERs that are stable across time (i.e. repeated in consecutive video frames) are selected as final character candidates.

A HOG-based descriptor is extracted for each of them and a classifier trained on standard Windows fonts predicts the per character probability. Given an English language model based on the probability of 4-grams, the system is able to choose the most likely combination of characters for each group of adjacent MSERs. The overall system is close to real-time (it can process a video at a relative speed of 0.7 times real-time).

Given the extracted words, a bag-of-word (sparse) descriptor is formed for each video. To decrease the descriptor sparsity, we include in the words' hypernyms (according to the Wordnet lexical database). We also found that including bi-words (i.e. pairs of words)

improves their distinctiveness and the overall performance. We train a linear SVM classifier to assign action scores to videos based on the text features.

Integration status

The KUL video OCR has not been integrated in AXES-V2.

AXES OPEN status

The component will not be part of AXES OPEN.

M5.6 Associated Text Normalization (NISV)

Objective

Since the format of metadata provided by the content owners is not identical in the first place, additional information present in these annotations has to be converted to a common schema. This converted metadata can then be utilized and indexed by other modules.

Description

NISV has worked on an XSL file enabling the conversion from their proprietary XML metadata format to this common Ebucore format. For every video of the dataset that will be used for the first experiment, NISV has provided a RDF file that has been processed off-line. Similar work will be performed for the datasets by other providers, with varying levels of integration into the processing chain.

Airbus DS has provided a WebLab service that enables to find the Ebucore file associated to a given document, to read its metadata and to add them to the document.

Integration status

The component that read Ebucore RDF files has been integrated.

AXES OPEN status

The component in charge of reading Ebucore RDF file will be part of AXES-V2.

M5.7 Entity Extraction (DCU)

Objective

In order to detect occurring entities like persons or places in textual content, this data is analyzed by the entity extraction module which performs a task similar to named entity recognition (NER). The detected entities are then used in the user interface, and for creating links between assets and segments.

Description

Named entity recognition task is implemented using components of the GATE framework. There are two components involved, including ANNIE plugin for German and English named entity recognition and Lingpipe for Dutch named entity recognition. The component assumes that all text is normalized prior to entry into the component.

ANNIE

ANNIE, as an Information Entity component, relies on a finite state algorithm to complete the named entity recognition task. ANNIE components are included in the GATE framework. AXES uses GATE directly to perform the named entity identification process by calling the ANNIE plugin. The components in ANNIE includes: Tokeniser, Gazetteer, Sentence Splitter, Semantic Tagger, Name Matcher and so on. The tokeniser splits the text into simple tokens such as numbers, punctuation and words based on different standards, such as in uppercase or lowercase. The rules of tokeniser include left hand side (LHS), which is a regular expression to match input text, and right hand side (RHS), describing the annotations to be added to annotation set. The gazetteer lists in ANNIE used are plain text files, with one entry per line. Each list represents a set of names, such as names of cities, organizations, days of the week, etc. The sentence splitter is a cascade of finite-state transducers which segments the text into sentences. ANNIE components are mainly meant to process English documents. Some simple model are also provided for German and French.

The German and English named entity recognition in AXES were developed by using Java GATE project. The jar file provides a full API including the ANNIE system. The program receives text document, which it assumes to be normalized. Based on an ANNIE gazetteer, plain text is split into tokens which are then marked as different entities. AXES entity extraction focuses on named entities such as location, personal name, and organization.

Lingpipe

The gazetteer provided by ANNIE in GATE does not contain functionality for processing Dutch. As a result, AXES uses another plugin, named Lingpipe, for Dutch named entity recognition. This solution is based on experimental work on work on Computational Natural Language Learning 2002 (ConNLL2002), from which Dutch training data for Lingpipe system is available. The Dutch named entity recognition solution involves two steps: Dutch training and entity extraction.

Different from ANNIE, which has a complete English gazetteer, Lingpipe doesn't have any gazetteer specially for Dutch. The training data used for AXES Dutch named entity recognition is from ConNLL2002. This conference provides two training data separately for Dutch and Spanish. In total there are four types of named entity defined in ConNLL2002, including person names (PER), organizations (ORG), locations (LOC) and miscellaneous names (MISC). In the training data, plain text has been processed with special tags to

identify their attributions. The format is B/I-PER/ORG/LOC/MISC. Here B and I means the beginning and ending of specific named entities, while the abbreviation following hyphen represents different named entity type.

There are three files in Dutch training data from ConNLL2002: one training data and two testing data. The current version of the AXES Dutch named entity recognition uses all three files as training data to expand gazetteer, since they follow the same format. The training process is performed offline.

During entity extraction step, the program receives language from other AXES packages. The GATE framework is used to distinguish languages based on input tags. Dutch is marked as "nt", while English is marked as "en". If Dutch is detected, Lingpipe Java code is called to process text according to Dutch gazetteer. The current version of Dutch named entity recognition focuses on three named entity types: personal name (marked as "B/I-PER"), organization name (marked as "B/I-ORG") and location (marked as "B/I-LOC").

Integration status

This component has been integrated in AXES system.

AXES OPEN status

The component as is will not be part of AXES OPEN; however an already integrated into the WebLab platform version of Gate NER service will be integrated into AXES Open.

CONCLUSION

This report describes the components developed in WP5 for the detection of entities in spoken and written text. Since both ASR and OCR are prerequisites for this detection, we provide components that can transform spoken words and on-screen text into textual representations. These are analysed by the entity extraction and used for indexing various datasets.

All components that are developed in WP5 are working in an off-line mode, i.e. they are processing the data during indexing time and do not perform any computations at query time.

REFERENCES

[1] Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... & Vesely, K. (2011, December). The Kaldi speech recognition toolkit. In *Proc. ASRU* (pp. 1-4).

[2] Federico, M., Bertoldi, N., & Cettolo, M. (2008, September). IRSTLM: an open source toolkit for handling large scale language models. In *Interspeech* (pp. 1618-1621).

[3] Stadtschnitzer, M., Schwenninger, J., Stein, D., & Koehler, J. Exploiting the large-scale German Broadcast Corpus to boost the Fraunhofer IAIS Speech Recognition System.

[4] Mikolov, T., Kombrink, S., Deoras, A., Burget, L., & Cernocky, J. (2011). RNNLM-Recurrent neural network language modeling toolkit. In *Proc. of the 2011 ASRU Workshop* (pp. 196-201).

[5] Baum, D., Schneider, D., Schwenninger, J., Samlowski, B., Winkler, T., & Köhler, J. (2010). DiSCo-a German evaluation corpus for challenging problems in the broadcast domain. *LREC 2010*.

[6] Bisani, M., & Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, *50*(5), 434-451.