INTERSPEECH 2019: Show & Tell Contribution
September 15–19, 2019, Graz, Austria

The SAIL LABS Media Mining Indexer and the CAVA Framework

Erinc Dikici, Gerhard Backfried, Jürgen Riedler

SAIL LABS Technology GmbH, Austria
{erinc.dikici,gerhard.backfried,juergen.riedler}@sail-labs.com

Abstract

In today's attention-driven news economy, rapid changes of topics and events go hand in hand with rapid changes of vocabulary and of language use. ASR systems aimed at transcribing contents pertaining to this fluid media landscape need to keep up-to-date in a continuous and dynamic manner. Static models, potentially created a long time ago, are hopelessly outdated within a short period of time. The frequent changes in vocabulary and wording need to be reflected in the models employed for optimal performance of transcription if one does not want to risk falling behind. In this demonstration paper we present the audio processing capabilities of the SAIL LABS Media Mining Indexer, and the CAVA Framework allowing semi-automatic and periodic updates of the ASR vocabulary and language model from relevant and new data.

Index Terms: automatic speech recognition, language model adaptation, keyword search

1. Introduction

Driven by an internationalized and attention-driven news economy, language encountered in the media nowadays evolves at a much faster pace than before. New words are being formed or adapted from existing words and borrowed from other languages, almost on a daily basis. Words and names appear instantly following a newsworthy event and often disappear as quickly as soon as attention has turned away and been directed to other issues. Some of these words appear and continue to "stay around" whereas others go out of use and silently disappear again. For example, the word *Brexit* did not exist a few years ago but by now has become one of the most popular words in the daily news, whereas the name *Edward Snowden* is hardly mentioned anymore. Such trends are often reversed or interrupted by the appearance of further facts, scandals or mere agenda-setting of interested actors. To equip ASR systems to be able to cope with such dynamic environments, a procedure to include out-of-vocabulary (OOV) words, drop outdated words and adjust the language models (LM) along the way seems to be required.

To this end, SAIL LABS' CAVA (Continuous Automatic Vocabulary and Language Model Adaptation) framework has been developed to mimic the principle of learning-and-forgetting by exploring relevant content from high-profile sources on the Web. The framework seeks out new (previously unknown) words along with their contexts, and updates/retrains the vocabulary and LM, respectively. These tasks are performed by a Web crawler, some NLP components, core algorithms for vocabulary selection, and the Language Model Toolkit (LMT), SAIL's tool for updating ASR models.

In this demonstration paper we introduce the audio processing functionality of the Media Mining Indexer, the LMT, and finally, the CAVA Framework.

2. Media Mining Indexer

The Media Mining Indexer (MMI) is a central component of the Media Mining System (MMS), SAIL LABS' state-of-the-art media monitoring and open-source intelligence platform. The MMI can process vast amounts of multimedia, auditory and textual data typically gathered from open sources (TV and radio stations, web pages and RSS feeds, several major social media platforms) in unstructured and semi-structured form. It can run as a standalone service on Windows as well as Linux operating systems and is also accessible through an API. The MMI typically operates in real-time and is self-aware of processing throughput, allowing it to adjust to changing environments automatically. The processing steps of the MMI comprise:

Audio pre-processing: The input video/audio is first extracted/converted into the native audio format (16kHz, 16-bit, mono, signed wave) followed by various normalization techniques. Feature extraction is applied subsequently to create sets of MFCCs and i-vectors at every centisecond.

Audio segmentation: The MMI uses a proprietary segmentation component, based on a speech/music/noise discriminator and phone-level decoding. Non-language sounds such as breathing and lip smacking are also accounted for and segments classified as containing sufficient amount of speech are passed on to ASR [1].

Speaker identification (SID): The SID component either determines a speaker's identity from a set of predefined target models, or the gender and age group of the speaker in case they are unknown. The set of target speakers can be extended or modified via the Speaker Identification Toolkit. Subsequent segments uttered by the same speaker are grouped together via clustering and marked as a speaker-turn [2].

Automatic speech recognition (ASR): The Kaldi-compatible ASR engine is designed for large vocabulary, speaker independent, multilingual real-time decoding of continuous speech. Besides from the 1-best output, lattice and N-best can also be generated with word- and utterance-based time tags and confidence scores. ASR is currently available for 25 languages: Albanian, Modern Standard Arabic, Egyptian Arabic, Levantine Arabic, Mandarin Chinese, Dutch, International English, US English, Farsi, French, German, Greek, Hebrew, Bahasa Indonesia, Italian, Bahasa Malaysia, Norwegian, Pashto, Polish, Romanian, Russian, Spanish, Mexican Spanish, Turkish, Urdu. Further languages are under development.

Post-processing: Following ASR, a Named Entity Detection (NED) module based on language-specific morphology, patterns and n-gram models tags the initial output text by various classes of named entities. These classes comprise persons, organizations and locations, but have also been extended to further classes such as diseases, disasters or financial terms. Subsequently Topic Classification (TC) categorizes segments according to a specific hierarchy of topics (inspired by Reuters). TC models are SVMs with linear kernels built from words [3]. Sentiment Analysis (Polarity-Detection) is applied as a further text-processing step, resulting in an XML-formatted output file [4]. This file is then uploaded together with a compressed version of the original media file onto the Media Mining Server (the big-data back-end of the MMS) where it can later be accessed for analysis, visualizations, search and retrieval purposes. In case of textual input, only the above text-processing components are invoked, resulting in a similar XML output file. In addition to the above listed languages for ASR an additional set of 8 languages is supported by the MMI for text-only processing: Catalan, Czech, Croatian, Bulgarian, Hungarian, Portuguese, Slovak and Swedish.

3. Language Model Toolkit (LMT)

The LMT allows one to update existing LMs and/or vocabularies as well as to build made-to-measure ones from scratch. It requires appropriate textual input to enable a sequence of processing steps yielding a model for ASR. Commonly, all text data undergo some pre-processing which includes language-dependent cleaning, normalization and tokenization, specific processing for numbers, compound-words, abbreviations, acronyms, spellings, etc. OOV words receive proper pronunciations by language-specific phonetic transcription tools (but may also be provided by end-users beforehand). User intervention is possible within each of the pre-processing stages. Finally as soon as the (re-)training is finished, a model for the MMI is created. This model is uploaded to a central repository and distributed to all MMIs within a cluster. The individual MMIs detect modified models automatically and re-load them in a seamless and transparent manner. The LMT can be run in an interactive manner from a GUI, or script-driven from a command line. The latter allows for integration into larger workflows.

4. CAVA Framework

CAVA - Continuous Vocabulary and Language Model Adaptation - provides a flexible and dynamic environment to keep models for ASR up-to-date. Based on a set of pre-defined high-level sources from the Web, a crawler component fetches textual content. This content is cleaned and pre-processed via sub-components of the LMT. CAVA uses an algorithm based on a variety of factors, such as frequency, recency and variety to determine updates of the vocabulary and of the LM. The processed web content, together with any other text files which the user may provide, are employed in combination with background information (the base model) for the adjustment of the LM. A brief overview of the CAVA Framework is shown in Figure 1.

CAVA is applied on a daily basis and refreshes the ASR models every night. Several measures are taken to evaluate newly created models before publishing them. Information about the changes to the vocabulary are communicated to selected members of the team. Should these changes be relevant from a "broader perspective", they can be made permanent. This functionality is also being used to connect changes and additions to the NED component with the ASR vocabulary, allowing to add new entries to the NED component and at the same time add them to the ASR models. During every cycle, the previous vocabulary changes are inspected which may result in the elimination of words (if not deemed relevant any longer). In this manner, ASR vocabularies grow and shrink over time, each day being adjusted to what is in the news at the current moment in time. The cycle can also be initiated on demand and could also be made dependent on the dynamics of news within a day. This way, more rapid fluctuations can be accommodated as well. CAVA thus allows for a flexible and (semi-)autonomous mode of operation and adaptation of systems deployed at SAIL LABS as well as at customer installations.

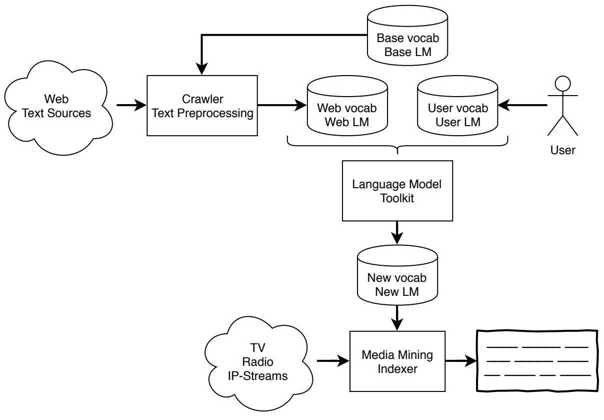

Figure 1: *Overview of the CAVA Framework.*

5. Conclusions and Outlook

The topics and events covered by the media and news industry change rapidly and in a continuous manner. This is reflected in the vocabulary and phrasing used to describe them. Technologies such as ASR, aiming to make contents searchable and manageable on a large scale, need to keep up to date and react to such changes. Not doing so risks falling behind and missing key terminology due to out-of-vocabulary effects. The CAVA Framework has been implemented at SAIL LABS, combining a series of components and technologies in order to semi-automatically update models for optimal performance. Human intervention is possible at various stages. The system has been in operation for several years, producing models on a daily (or, on-demand more frequent) basis for a set of languages. The combination with other components, such as the NED component, and re-transcription of previously transcribed data (with the updated vocabulary and LM) and the integration of data from Social Media form areas of activity in current CAVA development.

6. References

[1] D. Liu and F. Kubala, "Fast speaker change detection for broadcast news transcription and indexing," in *EUROSPEECH*, 1999.

[2] F. Kubala and D. Liu, "Online speaker clustering," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. I–333, 2004.

[3] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proceedings of the 10th European Conference on Machine Learning*, ser. ECML'98. Springer-Verlag, 1998, pp. 137–142.

[4] G. Shalunts, G. Backfried, and K. Prinz, "Sentiment analysis of German social media data for natural disasters," in *11th International Conference on Information Systems for Crisis Response and Management (ISCRAM)*, 2014.