Proceedings of the 1st Joint SLTU and CCURL Workshop (SLTU-CCURL 2020), pages 298–305
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020
© European Language Resources Association (ELRA), licensed under CC-BY-NC

Adapting Language Specific Components of Cross-Media Analysis Frameworks to Less-Resourced Languages: the Case of Amharic

Yonas Woldemariam, Adam Dahlgren
Dept. Computing Science, Umeå University, Sweden
{yonasd, dali}@cs.umu.se

Abstract

We present an ASR based pipeline for Amharic that orchestrates NLP components within a cross media analysis framework (CMAF). One of the major challenges that are inherently associated with CMAFs is effectively addressing multi-lingual issues. As a result, many languages remain under-resourced and fail to leverage out of available media analysis solutions. Although spoken natively by over 22 million people and there is an ever-increasing amount of Amharic multimedia content on the Web, querying them with simple text search is difficult. Searching for, especially audio/video content with simple key words, is even hard as they exist in their raw form. In this study, we introduce a spoken and textual content processing workflow into a CMAF for Amharic. We design an ASR-named entity recognition (NER) pipeline that includes three main components: ASR, a transliterator and NER. We explore various acoustic modeling techniques and develop an OpenNLP-based NER extractor along with a transliterator that interfaces between ASR and NER. The designed ASR-NER pipeline for Amharic promotes the multi-lingual support of CMAFs. Also, the state-of-the art design principles and techniques employed in this study shed light for other less-resourced languages, particularly the Semitic ones.

Keywords: Less-resourced, Amharic, Cross-media analysis, Speech recognition, named entity recognition

1. Introduction and Background

Automatic Speech Recognition (ASR) and Named Entity Recognition (NER) perform information extraction tasks on spoken and textual documents respectively. ASR generates a transcription text from speech data. ASR technologies have been used for many applications such as spoken document indexing and retrieval (Chang et al., 2005; Aichroth et al., 2015; Le et al., 2017), spoken dialogue systems (Ivanov et al., 2015), speech translation (Stüker et al., 2011), and so on. NER is used to identify and extract entity mentions, such as names of people, locations, etc from textual contents. In natural text analysis, NER performs a pre-processing task for downstream annotators (e.g., syntactic parsers (Marneffe et al., 2006)) and identifies proper nouns and classifies them into known categories (e.g., Person, Place, Organization and so on). While both ASR and NER are essential to solve specific problems in isolation, it is also possible to join them systematically to operate on same media (e.g., a webpage containing text and audio tracks), and apply them in succession to add contextual information on the metadata associated with audio/video contents for semantic search (Aichroth et al., 2015; Le et al., 2017).

Depending on the purpose of the application in question, ASR and NER can be combined in various ways. For example, in cross-media analysis frameworks such as EUMSSI[1] (Event Understanding through Multimodal Social Stream Interpretation) and MICO[2] (Media in Context), their combination is defined as an analysis workflow or analysis-chain called an ASR-NER pipeline that basically includes speech transcription and named entity extraction services. Within these frameworks there also exist complex multimedia analysis pipelines designed to meet the requirements of complex information retrieval use cases, for instance, searching for video shots, where a person (in the shots) says something about a specific political issue using a keywords-driven approach.

Nowadays, there are plenty of multimedia extraction tools used to make searching web contents convenient. However, most of these tools are developed for well researched and resourced languages such as English and Spanish, and specific domains of applications. Due to this reason, many languages including Amharic, remained under-sourced. That severely limits the access of information available in those languages. There are some studies (Abate et al., 2009; Yifiru, 2003; Belay, 2014; Demeke and Hailemariam, 2012) and contributions on building language technologies for Amharic, but most of them are developed as proof-of-concept prototypes with very limited data and resources(Gauthier et al., 2016; ELRA-W0074, 2014; HaBiT, 2016). As a result, it is often challenging to get computational linguistic resources for Amharic required for either NLP studies or commercial use.

Amharic is the official language of Ethiopia, spoken by over 22 million people, also according to the latest census carried out by Central Statistical Agency of Ethiopia[3], the second most spoken Semitic language next to Arabic. The writing system of Amharic is called "fidel"; shared with the other Semitic language of Ethiopia, Tigrinya. Amharic has a unique writing system and its basic alphabet units have a consonant-vowel (CV) syllabic structure, usually vowels are omitted in the written form of CV. There is an ever-increasing amount of Amharic digital contents of various types: text, images, audio, video, etc. on the Web due to emerging information sharing platforms such as social media and video hosting sites. However, querying them with simple text search is difficult, especially audio and video contents, is even very hard as they exist in raw format (not well indexed). Thus, obviously it is very demanding to have

[1] https://www.eumssi.eu
[2] https://www.mico-project.eu
[3] https://www.csa.gov.et

linguistically motivated multimedia analysis and extraction tools that could potentially deal with language-related concerns and make Amharic contents more searchable through keywords.

The most reasonable and affordable solution is to use open-source multi-lingual information extraction frameworks that provide media analysis, extraction and indexing, search and retrieval services, though they require language models of certain types. One of existing open-source media analysis solutions, is the MICO platform, though it is at early stage of its release. Ideally, the platform allows extraction of multimedia contents of different languages using the corresponding language models. Within the platform, there are a number of pre-defined analysis pipelines along with their metadata extractors.

The aim of this study is to investigate adapting language specific components of MICO for Amharic. That could potentially be extended to other languages, particularly Semitic ones as they share similar orthography (e.g., Tigrinya) and phonology (e.g., Arabic). Within MICO, there are several natural language dependent multimedia analysis components such as text classification and text language detection including the ASR-NER pipeline. However, we only focus on designing of an ASR-NER pipeline for Amharic using the design principles, the standards and the technologies used in MICO. The pipeline could be considered as the first step to be able to use the MICO platform and for developing other important metadata extractors to analyze Amharic contents. Indeed, the pipeline is useful in itself, at least to index video /audio contents with extracted entities. To completely benefit from the platform more effort is needed in the direction of identifying and adapting other language dependent analysis components, for instance, sentiment analysis. We basically develop Kaldi-based acoustic models, a transliterator and an OpenNLP based NER extractor, to build the **Amharic ASR-NER pipeline**.

We got motivated for this study as we are one the partners of the MICO project and responsible for implementing NLP tools. While most of the implementation is done only for English, the MICO architecture allows for the integration of other language models via its API. Nevertheless, it is challenging to adapt MICO to under-resourced languages due to its requirement of trained language models that strictly satisfy the underlying design principles. This presents an opportunity to investigate the possibilities of adapting relevant language models for Amharic.

We discuss related works in Section 2., the MICO platform in Section 3., the designed ASR-NER pipeline and the discussion in Section 4., the challenges and solutions in Section 5. and, finally, future work and conclusion in Section 6..

2. Related Work

There are a number of papers (Magnini et al., 2013; Hori and Nakamura, 2006) on extraction of named entities on speech transcripts on digital spoken archives for various purposes, though it is hardly possible to get any for Amharic. There are also a few research projects that investigated the introduction of an ASR-NER pipeline in multi-modal cross-media analysis frameworks for different types of languages. We primarily focus on discussing the methods used and the results achieved by these projects, as they probably best put our study into perspective, namely MICO and EUMSSI. In addition to that, although there is no published literature on the task of NER on speech transcription for Amharic, we present a brief review of research works on standalone speech recognition and named entity recognition conducted independently from each other.

During the development of MICO metadata extractors, special attention was given to the ASR component due to the fact that most extractors, particularly text analysis components including NER heavily depend on the result produced by the ASR extractor. In order to achieve high-quality speech transcription, state-of-the-art open-source and proprietary libraries for ASR have been well studied and evaluated against sample video contents, then the respective comparative analysis was carried out beforehand. Consequently, Kaldi[4] was chosen based on the criterion of accuracy and other technical reasons. The other good quality of Kaldi is its multi-lingual support. Most of the experiments that make use of Kaldi within MICO were effectively carried out only for English, though the MICO Showcases were planned for Arabic and Italian as well. The most challenging part of training Kaldi is that preparing a parallel corpus (speech and text) is quite costly.

Within MICO, the ASR is implemented as a speech-to-text pipeline to analyze video content and produce the corresponding text transcription in various formats. The pipeline includes audio-demultiplexing, for extracting and down-sampling the audio signal from video content, speaker diarization (Meignier and Merlin, 2010; Tranter and Reynolds, 2006) for segmenting information along with gender classification and speaker partitioning, speech transcription, for transcribing audio signal into text. The resulting textual content generated from the pipeline is further analyzed by text analysis components including the NER extractor.

The NER extractor provides a named entity extraction service on-demand when requested by other registered extractors requiring (depending) on the output produced by it. NER also takes plain text (with a text/plain MIME type) from other possible sources of textual contents such as forum discussion posts after pre-processing and parsing tasks. The NER extractor is based on the OpenNLP toolkit, that is an open-source library providing a NER service. MICO provides OpenNLP-based NER language models for English, German, Spanish and Italian, and allows an integration of models for other languages.

The ASR-NER pipeline introduced in MICO performs analysis workflows, for instance, detecting a person in a video, by collaborating with image analysis components such as the face detection extractor. Some preliminary showcases have been demonstrated by the use case partners, for instance, InsideOut10 (one of the use case partners of the MICO project) built a showcase application that retrieves video shots containing a specific person talking about a specific title (Kurz et al., 2015).

[4] http://kaldi-asr.org

The EUMSSI platform basically provides multimodal analytics and interpretation services for different types of data obtained from various online media sources. (their demo is available on[5]). EUMSSI seems to mainly target journalists as end users, automating their time-consuming tasks of organizing information about various events from different online and traditional data sources providing un/structured contents. The platform allows to search multimedia contents aggregated and filtered from media search engines in an interactive fashion, then enriching, contextualizing the media with extracted metadata and retrieves the result with the multimodal approach.

The NER component of EUMSSI is based on the Stanford NER (Finkel et al., 2005), running on the transcription generated by ASR and text extracted by OCR (Optical character recognition) from video contents, in addition to other types of textual contents from news and social media. The transcription returned from the ASR service is normalized by an auxiliary component beforehand. The ASR-NER pipeline implemented in EUMSSI, is used to annotate the speech segments uttered by each speaker shown in a video with the corresponding transcriptions and mentioned names. The resulting information is intended to get combined with the annotations obtained from the face recognition component, that enables video retrieval applications to support different search options, for instance, retrieving quotations of peoples (Le et al., 2017).

There are also several studies on named entity extraction on speech transcripts for independent NLP systems or audio/video analysis frameworks. For example, in the Evalita (evaluation campaign of NLP and Speech tools for Italian) 2011 workshop (Magnini et al., 2013), one of the tasks was named entity recognition on transcribed broadcast news. The purpose is to investigate the impact of the transcription errors on NLP systems and explore NER approaches that cope with the peculiarities of the resulting transcripts from ASR systems.

There are a number of studies on the design and development of ASR and NER systems for Amharic. Relatively, NER is a less researched area than ASR. The survey in (Abate et al., 2009), summarizes ASR research works attempted for Amharic, ranging from syllable to sentence level detection, from speaker dependent to speaker independent speech recognition. According to the survey most of the works are done using quite similar techniques i.e. HMM (Hidden Markov Model) (Rabiner, 1989) and tools such as HTK (HMM Tool Kit). There is an attempt to develop and integrate an ASR system into the Microsoft Word application to enable it to receive file related commands. The survey also pointed out that the major reasons, why the ASR systems failed to be used in speech applications, to mentions some of them: they are trained on read speech with a limited dataset and fail to handle germination and morphological variations. There are also a few unpublished research works on Amharic NER (Mehamed, 2010; Belay, 2014). The recent work (Gambäck and Sikdar, 2017) introduced deep learning with the skip-gram word-embedding technique by extending the previous works. The authors in (Gambäck and Sikdar, 2017) developed Amharic NER prototypes using the same method i.e., Conditional Random Fields (Sobhana et al., 2010; Finkel et al., 2005) and the same corpus as in (Mehamed, 2010; Belay, 2014) but different subsets, and obtained different results.

[5]http://demo.eumssi.eu/demo/

3. The MICO Platform

Basically the MICO plaform provides media analysis, metadata publishing, search and recommendation services (described in (Aichroth et al., 2015)). It has three types of metadata extractors, textual extractors for performing linguistic analysis such as parsing, sentiment analysis and text classification, image extractors for performing image analysis for detecting and human faces and animals from images, audio extractors for performing different speech analysis tasks such as detecting whether audio signals contain music or speech, and extracting audio tracks from video content and producing a transcription.

Metadata extractors interact and collaborate with each other in automatic fashion via a service orchestration component (aka broker) to put a media resource in context. Several semantic web technologies such as Apache Marmotta[6] and SPARQL-MM[7] are used for storing the metadata annotation of analysis results in a RDF format and querying the metadata respectively. The Apache Hadoop[8] distributed file system is used for binary data, and Apache Solr[9] for full-text searching.

4. The Amharic ASR-NER Pipeline

The Amharic ASR-NER pipeline designed in this study includes three main components: ASR, a transliterator and NER (see Figure 2). The pipeline performs extracting named mentioned from audio and video contents. Within the MICO architecture, the core ASR component needs to be connected with pre-processing and post-processing components, that forms a speech-to-text sub-pipeline.

There are two pre-processing components, namely audidemux and LIUM diarization. The former does extracting audio tracks from a video input and down-sampling the audio tracks. The later does segmenting the audio tracks into smaller units using gender and speaker information. The post-processing component, namely XML2text transforms the output file (in the text/xml format) generated by the core ASR component to plain text (text/plain) required by the NER component.

4.1. Building the Amharic ASR

We explored and applied three different acoustic modeling techniques, namely GMM-HMMs (Gaussian mixture model-hidden Markov model), DNN (Deep Neural Networks) and SGMM-HMMs (Subspace Gaussian Mixture model) to build the Amharic ASR. The Kaldi (Povey et al., 2011b) framework is used as an open speech recognition toolkit. While DNN-HMM is the state-of-the-art ASR modeling technique, SGMM-HMM (Povey et al., 2011a)

[6]http://marmotta.apache.org

[7]http://marmotta.apache.org/kiwi/sparql-mm.html

[8]http://hadoop.apache.org

[9]http://lucene.apache.org/solr/

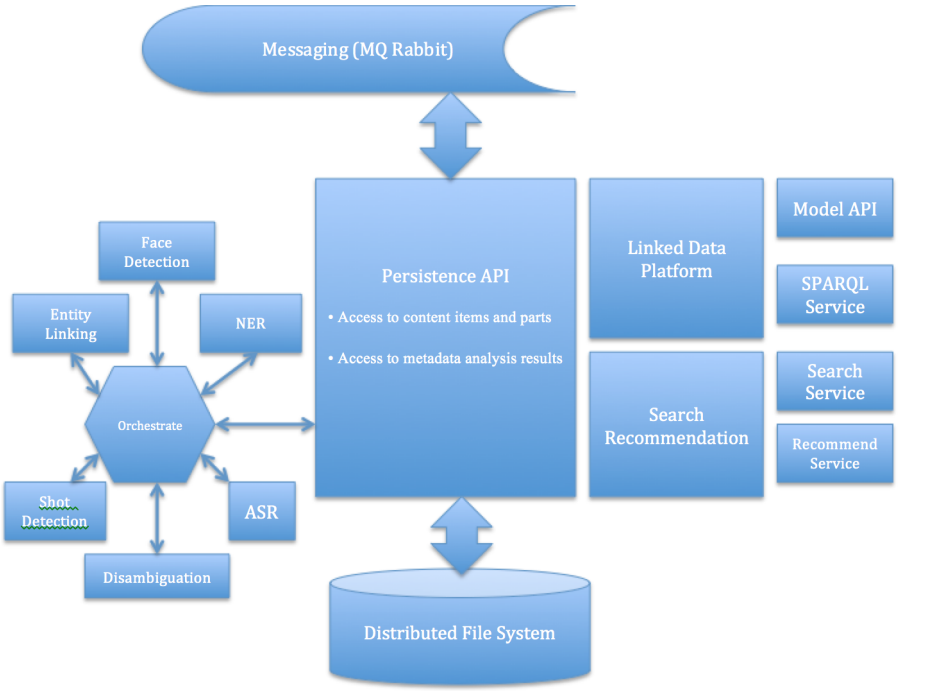

Figure 1: The Architecture of the MICO Platform

is an extension of GMM-HMM that is one of conventional acoustic modeling approaches. In HMM-based ASR systems, (S)GMMs and DNN estimate probability distributions of phonemes over HMM states given observations (acoustic inputs). During training they compute model parameters (e.g, mean vectors and covariance matrices) from training data, (S)GMMs use the expectation maximization algorithm, whereas DNN uses stochastic gradient descent and back-propagation to adjust weights and biases.

As a result, three acoustic models have been built using each technique with a parallel speech-transcription corpus (Gauthier et al., 2016), a pronunciation lexicon and a language model. Originally, the raw corpus was prepared for the study in (Tachbelie et al., 2014), it is about 20 and 2 hours of speech for training and testing respectively. We built a 5-gram language model using the SRILM[10] language modeling toolkit with the Kneser-Ney smoothing method.

All the three acoustic models are trained with 13 Mel-frequency cepstrum coefficients (MFCCs) features, followed by linear discriminant analysis (LDA) and transformation, maximum likelihood transform (MLLT). Also, feature-space maximum likelihood linear regression (fMLLR) has been used as a speaker adaptation technique. The models are evaluated on the same test set containing 6203 words using the Word Error Rate (WER) metric, and obtained a WER of 50.88%, 38.72%, and 46.25% for GMM-HMM, DNN-HMM, SGMM-HMM respectively.

4.1.1. Discussion of ASR Acoustic Models

The experimental results obtained from the ASR models evaluation show that the DNN-HMM model outperforms than GMM-HMM and SGMM-HMM models, with a WER of 12.16% and 7.53% respectively. The SGMM-HMM model in turn outperforms GMM-HMM with a WER of 4.63%. In our experiments, the GMM-HMM acoustic model gets trained with utterance-level transcriptions, the resulting model is used to generate phone alignments for DNN training. For that reason, the DNN acoustic model appears to have the best performance (regarding WER). DNNs also have the ability to capture larger context (larger window of frames), for example, the DNN in this study, is trained with 5 preceding and 5 following frames. Moreover, the number of model parameters (weights) computed by DNN is extremely larger than (S)GMMs, that potentially help learn the complex relationship between acoustic features extracted from input speech signal and their associated sequence of phonemes. For (S)GMMs, the training data seems to be too small to effectively model the distributions of acoustic units and generalize for new input data. Compared with state-of-the-art ASR systems built for other languages (Wang and Zheng, 2015; Xiong et al., 2018; Ghahremani et al., 2017), for instance, authors in (Xiong et al., 2018) achieved a 5.1% of WER, that suggests more tasks are needed to improve our ASR. Unlike these studies where a large amount of data is used to train acoustic models, in our study the amount of training data is limited to 20 hours. Basically, the results obtained in this study could be improved by increasing the size of the training data, including a large vocabulary to deal with the problem of out-of-vocabulary (OOV) and language models with different size of n-gram (e.g., (n=3 to 7)). However, preparing such resources is quite expensive and time-consuming, especially for less studied and under-resourced languages like Amharic. Therefore, adapting from pre-trained acous-

[10]http://www.speech.sri.com/projects/srilm/

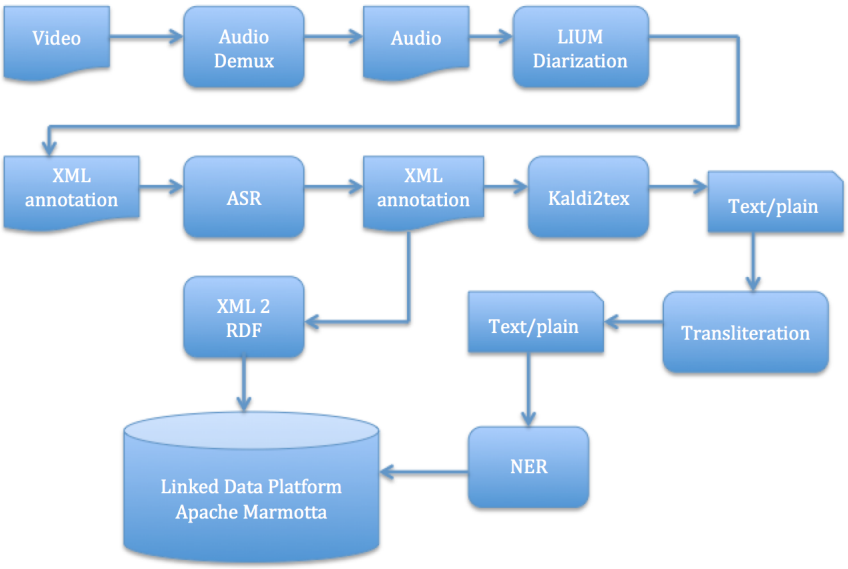

Figure 2: The Amharic ASR-NER Pipeline within a Cross-media Analysis Framework

tic models trained on other languages, particularly well resourced ones, seems to be more reasonable. Besides, multilingual model training (Ghoshal et al., 2013; Wang and Zheng, 2015; Feng and Lee, 2018) could be considered, where under multiple languages (including less-resourced ones) get trained together. Then, the resulting acoustic model can be used to produce speech transcriptions of any of these languages. However, it, in turn, requires a huge amount of multilingual resources including parallel speech-text corpora, language models and pronunciation dictionaries (Besacier et al., 2014; Wang and Zheng, 2015). The languages also need to be related and share the same phone set. In practice, it is too difficult to meet these requirements, especially the problem gets worse when it comes to Amharic and other Semitic languages as they are not yet studied using this approach. The other alternative approach is transfer learning (Wang and Zheng, 2015; Huang et al., 2013; Ghahremani et al., 2017), that allows an acoustic model trained on one language to get adapted to other languages. That is possible via sharing parameters learned during neural net based model training of one language to others.

4.2. The Amharic NER Extractor and Transliterator

Given the very limited alternative choices for Amharic NER models, we used the NER model developed as part of the master thesis by Belay M. (Belay, 2014). As the original model was built in a format which is incompatible with other meta-data extractors within MICO, the data needs to be re-labeled manually to train the OpenNLP name finder models. That is the right format supported by MICO. However, it was possible only to label the small portion of the whole data used in (Belay, 2014) with only the following entities: persons, locations and organizations. The models are trained using machine learning algorithms provided by OpenNLP: MaxEntropy (Berger et al., 1996) and Perceptron (Kazama and Torisawa, 2007). As shown in **Table 1**, the Perceptron based model outperforms the MaxEntropy based model, regarding all considered metrics. As both the training and testing sets are quite small (compared to the standard requirement, i.e., 15K sentences, but here the models are trained on 420 sentences and evaluated on 45 sentences along with 126 entities) for generalization, the evaluation details are not included in this study. In order to use the models, we then developed a Java-based application that loads the NER models and extracts named mentions from speech transcriptions.

While the NER models are trained on the transliterated form of Amharic text, the ASR acoustic models are trained on transcripts with the actual Amharic orthography. Because it seems to be most open-source NLP research tools are primarily designed for English, Amharic NLP studies tend to use an Amharic-English transliteration scheme (Sebsibe et al., 2004) in their prototype development. In order to support the interfacing of ASR with NER, we implemented a simple rule-based transliteration program that converts Amharic scripts to its corresponding English transliteration form.

5. Challenges and Solutions

Since the main goal of this research is to make less-resourced languages beneficial out of media analysis technologies built for resource rich languages, by dealing with issues related with scarcity of computational linguistic resources, most of the challenges faced in the course of the study is inherently associated with the lack of resources. In addition, we assumed that the resources that have been available can be modified with reasonable amount of configuration tasks and then would fit to the designed experimental settings, but a number of evaluations (compatibility tests) have shown that they turned out to require to get

Classifier	Recognized Entity	Metric (in%)		
		Precision	Recall	F-Score
MaxEntropy	Person	79.17	73.08	76.00
	Organization	84.93	71.26	77.50
	Location	16.67	7.69	10.53
Perceptron	Person	**85.71**	**92.31**	**88.89**
	Organization	64.15	78.16	70.47
	Location	66.67	15.38	25.00

Table 1: NER Models Evaluation Results

transformed with much amount of works. For example, re-labeling the NER dataset, improving the quality of the acoustic models and so on.

As part of our study, we also observed major important issues that arise from the natural language perspective during the adaptation of MICO for Amharic. The issues are very important for other new languages as well to be considered in advance. That mostly include availability of compatible language dependent analysis components and other pre/post processing auxiliary utilities (e.g., language detection, file format adaptors). In order to effectively meet the compatibility requirements (e.g., data models, file formats), one needs to closely look at the synergies and the dependencies between all meta-data extractors.

Although MICO aims to provide an open data model via its API, at the current stage of its implementation new languages are required to strictly adhere some specifications, for example, while NER models need to be in an OpenNLP based, ASR acoustic models in deep neural net. Among other important language specific components ASR and NER seem to be very foundational and take a high priority, as others downstream extractors such as sentiment analysis, text classification and topic detection rely on the quality provided by the ASR-NER chain.

The other problem is related with computational resources, training the DNN-HMM model has been challenging due to the requirement of GPU processors along with the queue scheduling service configuration. Although it is extremely slow, the training has been done on our CPU machine with a slight job-scheduling configuration task.

Lastly, it concerns the interfacing Amharic ASR with NER. The transcription generated by ASR is in the actual orthographic form of Amharic, where as the NER models are trained on an English-transliteration form. Thus, to support the NER models a simple rule-based transliteration program has been written.

6. Conclusions and Future Work

We identified language dependent analysis components that are viewed as a high priority including ASR and NER, within a cross-media analysis platform. We designed an ASR-NER analysis pipeline for Amharic based on state-of-the art design principles and techniques employed in cross-media solutions, thus promoting the multi-lingual support of the MICO platform. Moreover, this study provides a chance to further explore ASR methods introduced to potentially support under-resourced languages such as transfer learning. Moreover,the quality of both the ASR and NER models can be enhanced with availability of more data and improve the transliteration phase to reasonable quality in the future. Also, as this study has been done during the early release stages of the MICO platform for English, it has been hard to fully support Amharic for more detailed experiments. However, for future it would be interesting to carry out additional evaluations across other parts of the pipeline. Generally, other languages somehow take advantages of the methods proposed here, especially those that share a similar orthographic structure with Amharic, such as Tigrinya. Also, the method can be easily extended for other Semitic languages such as Arabic and Hebrew.

7. Acknowledgments

We acknowledge the financial support from the EU FP7 MICO project. We also thank Mikyas Belay for providing Amharic NER models used in this study.

8. Bibliographical References

Abate, S., Tachbelie, M., and Menze, W. (2009). Amharic speech recognition: Past, present and future. In *Proceedings of the 16th International Conference of Ethiopian Studies*, pages 1391–1401.

Aichroth, P., Weigel, C., Kurz, T., Stadler, H., Drewes, F., Björklund, J., Schlegel, K., Berndl, E., Perez, A., Bowyer, A., and Volpini, A. (2015). Mico-media in context. In *Proceedings of 2015 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*, pages 1–4.

Belay, M. (2014). Amharic Named Entity Recognition Using a Hybrid Approach. Master's thesis, School of Information Informatics, Addis Ababa University.

Berger, A. L., Pietra, S. D., and Pietra, V. J. D. (1996). A maximum entropy approach to natural language processing. *Computational Linguistics*, 22:39–71.

Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages. *Speech Communication*, 56:85–100.

Chang, S.-F., Manmatha, R., and Chua, T.-S. (2005). Combining text and audio-visible features in video indexing. In *Acoustics, Speech, and Signal Processing*, pages 1005–1008.

Demeke, Y. and Hailemariam, S. (2012). Duration modeling of phonemes for amharic text to speech system. In *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*, pages 1–7.

Feng, S. and Lee, T. (2018). Improving cross-lingual knowledge transferability using multilingual tdnn-blstm with language-dependent pre-final layer. In *Proceedings of Interspeech*, pages 2439–2443.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370.

Gambäck, B. and Sikdar, U. K. (2017). Named entity recognition for amharic using deep learning. In *2017 IST-Africa Week Conference (IST-Africa)*, pages 1–8.

Ghahremani, P., Manohar, V., Hadian, H., Povey, D., and Khudanpur, S. (2017). Investigation of transfer learning for asr using lf-mmi trained neural networks. In *Proceedings of 2017 IEEE Automatic Speech Recognition and Understanding Workshop*, pages 279–286.

Ghoshal, A., Swietojanski, P., and Renals, S. (2013). Multilingual training of deep neural networks. In *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7319–7323.

Hori, T. and Nakamura, A. (2006). An extremely large vocabulary approach to named entity extraction from speech. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, pages 973–976.

Huang, J.-T., Li, J., Yu, D., Deng, L., and Gong, Y. (2013). Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *Proceedings of 2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7304–7308.

Ivanov, A. V., Ramanarayanan, V., Suendermann-Oeft, D., Lopez, M., Evanini, K., and Tau, J. (2015). Automated speech recognition technology for dialogue interaction with non-native interlocutors. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 134–138.

Kazama, J. and Torisawa, K. (2007). A new perceptron algorithm for sequence labeling with non-local features. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 315–324.

Kurz, T., Schlegel, K., and Kosch, H. (2015). Enabling access to linked media with sparql-mm. In *Proceedings of the 24th International Conference on World Wide Web*.

Le, N., Bredin, H., Sargent, G., India, M., Lopez-Otero, P., Barras, C., Guinaudeau, C., Gravier, G., da Fonseca, G. B., Freire, I. L., do Patrocínio, Z. K. G., Guimarães, S. J. F., Martí, G., Morros, J. R., Hernando, J., Fernández, L. D., García-Mateo, C., Meignier, S., and Odobez, J.-M. (2017). Towards large scale multimedia indexing: A case study on person discovery in broadcast news. In *Proceedings of International Workshop on Content-Based Multimedia Retrieval*, pages 1–6.

Magnini, B., Cutugno, F., Falcone, M., and Pianta, E. (2013). Evaluation of natural language and speech tools for italian. In *Lecture Notes in Computer Science*, pages 98–106.

Marneffe, M.-C., MacCartney, B., and Manning, C. (2006). Generating typed dependency parses from phrase structure parses. In *InProc. 5th Interna-tional Conference on Language Resources and Evaluation (LREC 2006)*, pages 449–454.

Mehamed, M. (2010). Amharic Named Entity Recognition. Master's thesis, College of Natural Sciences, Addis Ababa University.

Meignier, S. and Merlin, T. (2010). Lium spkdiarization: An open source toolkit for diarization. In *CMU SPUD Workshop*.

Povey, D., Burget, L., Agarwal, M., Akyazi, P., Feng, K., Ghoshal, A., Glembek, O., Goel, N. K., Karafiát, M., Rastrow, A., Rose, R. C., Schwarz, P., and Thomas, S. (2011a). The subspace gaussian mixture model - a structured model for speech recognition. *Computer Speech and Language*, 25(2):404–439.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., and Vesely, K. (2011b). The Kaldi speech recognition toolkit. In *Proceedings of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Sebsibe, H., Prahallad, K., Alan, B., Rohit, K., and Rajeev, S. (2004). Unit selection voice for amharic using festvox. In *Fifth ISCA Workshop on Speech Synthesis*, pages 103–107.

Sobhana, N., Mitra, P., and Ghosh, S. (2010). Conditional random field based named entity recognition in geological text. *International Journal of Computer Applications*, 1(3):143–147.

Stüker, S., Kilgour, K., and Niehues, J. (2011). Quaero speech-to-text and text translation evaluation systems. In *High Performance Computing in Science and Engineering'10*, pages 529–542. Springer.

Tachbelie, M., Abate, S., and Besacier, L. (2014). Using different acoustic, lexical and language modeling units for asr of an under-resourced language - amharic. *Speech Communication*, 56:181–194.

Tranter, S. and Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. *IEEE Transactions on Audio, Speech, and Language Processing*, 14:1557–1565.

Wang, D. and Zheng, T. F. (2015). Transfer learning for speech and language processing. In *Proceedings of 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 1225–1237.

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., and Stolcke, A. (2018). The microsoft 2017 conversational speech recognition system. In *Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5934–5938.

Yifiru, M. (2003). Automatic Amharic Speech Recognition System to Command and Control Computers. Mas-

ter's thesis, School of Information Studies for Africa, Addis Ababa University.

9. Language Resource References

ELRA-W0074. (2014). *Amharic-English bilingual corpus, distributed via ELRA,1.0*. distributed via ELRA,1.0, 1.0, ISLRN 590-255-335-719-0.

Elodie Gauthier and Laurent Besacier and Sylvie Voisin and Michael Melese and Uriel Pascal Elingui. (2016). *ALFFA (African Languages in the Field: speech Fundamentals and Automation)*. European Language Resources Association (ELRA), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), ISLRN SLR25.

HaBiT. (2016). *Harvesting big text data for under-resourced languages*. distributed via Natural Language Processing Centre, Faculty of Informatics, Masaryk University.