Panorama: BBN Participation in SM-KBP 2019

Sancar Adali, Roger Bock, Daniel Ellard, Joshua Fasching, John Greve, Ilana Heintz, Kevin Jett, Zhuolin Jiang, Clay Riley, Alex Zamanian, Le Zhang
Raytheon BBN Technologies
Cambridge, MA 02138, USA

Abstract

We describe the BBN submission to the TAC 2019 Streaming Multimedia Knowledge Base Population (SM-KBP) track. In a pipeline similar to that of our 2018 submission with enhancements to usability and several analytic modules, we processed multimedia data to create coherent knowledge elements (entities, relations, and events) in NIST-restricted AIDA Interchange Format.

1 Introduction

The Linguistic Data Consortium (LDC) provided two thousand documents with information regarding relationships of parent and child documents, the file type of child documents, and additional metadata. Our goal was to provide, for each parent document, a set of entities, relations, and events (knowledge elements, or KEs) aligned with the AIDA ontology, such that a single real world instance corresponded to a single KE for the parent document, with justifications reaching back into all child documents. For example, in a parent document that includes text, an image, and a video, we produce a single Entity for Person A, and supply TextJustifications resulting from information extraction over the text, ImageJustifications resulting from a FaceID application run over the image, and KeyFrameVideoJustifications for KEs found in the transcribed speech and extracted frames of the video. Each parent document is then associated with a mini-knowledge base representing all KEs extracted from all media types found in the child documents.

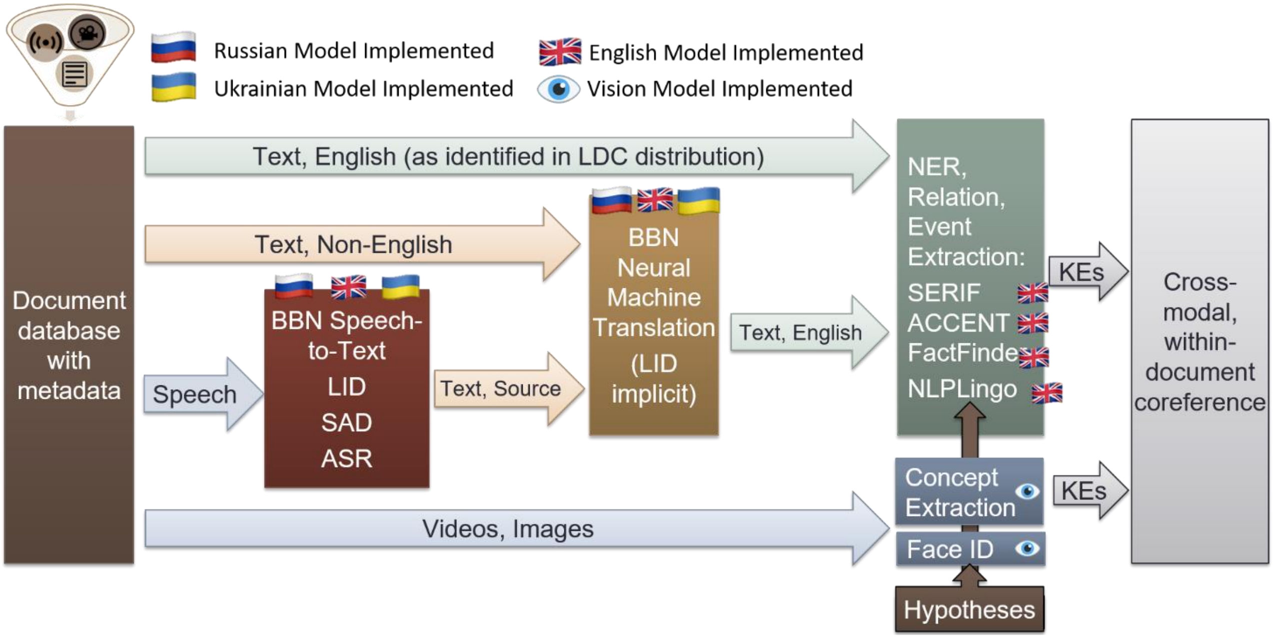
Russian Model Implemented
Ukrainian Model Implemented
English Model Implemented
Vision Model Implemented
Document database with metadata
Text, English (as identified in LDC distribution)
Text, Non-English
Speech
BBN Speech-to-Text
LID
SAD
ASR
Text, Source
BBN Neural Machine Translation (LID implicit)
Text, English
NER, Relation, Event Extraction:
SERIF
ACCENT
FactFinder
NLPLingo
KEs
Concept Extraction
Face ID
Hypotheses
Videos, Images
KEs
Cross-modal, within-document coreference

Figure 1: Panorama Pipeline Components and Workflows. LID: Language ID; SAD: Speech Activity Detection; ASR: Automatic Speech Recognition; NER: Named Entity Recognition; LDC: Linguistic Data Consortium; KE: Knowledge Element

2 Panorama Pipeline

Figure 1 depicts the Panorama pipeline, including all individual components and possible workflows. As compared to last year, we have removed the OCR component, due to lack of useful outcomes. Also different from last year is an overall improvement in running the pipeline. With a single command, our pipeline runs all processes and analytics, in order, over the correct files, with no manual message-passing to locate intermediate outcomes. Parallel processing within and across analytics is handled by a wrapper around Sun Grid Engine processes which handles job execution graphs and load balancing. Message passing is handled by access to a sqlite database that stores the paths to all raw data paths, all intermediate outcomes (e.g. speech transcription and machine translated text), and all final outcomes.

3 Speech-to-Text

Our Speech-To-Text component runs three separate steps: speech activity detection (SAD), language identification (LID), and automatic speech recognition (ASR).

The audio tracks extracted from video may contain long segments of music or other non-speech audio signal. In order to reduce false positives that may result from running ASR over non-speech data, we first run speech activity detection to extract pairs of timestamps where human speech is found. For this task we apply an existing, standardized SAD model.

We train a LID model specifically for the languages English, Russian, and Ukrainian, using data from the AIDA background corpus. For English and Russian, we train LID using only audio segments that are aligned to transcripts. For Ukrainian, we were not able to use the transcripts, and used all of the raw audio. As a result, as shown in Table 1, the LID accuracy on Ukrainian is lower than that of Russian and English.

Table 1: LID Accuracy on held-out AIDA audio data

Language	LID Accuracy
English	1.0
Russian	0.914
Ukrainian	0.346

At both training and test time, we made the assumption that each audio file contained a single language, and transcribed all segments with the language model for the highest-scoring language resulting from LID.

As for speech transcription, we have independent models for English, Russian, and Ukrainian. In the last case, we have updated our model from the previous evaluation by training a new acoustic model using the additional Ukrainian broadcast news data released in 2019, and adding additional data to the language model from the AIDA background corpus. In Table 2Table 3, we report the overall word error rate as well as Mean Average Precision (MAP) for in-vocabulary (iv), out-of-vocabulary (oov), and infrequent (rare) terms for each language. We also measure the MAP for target words (targ) that are scenario-relevant in Russian and Ukrainian.

Table 2: Word Error Rate (WER) and Mean Average Precision (MAP) on held-out AIDA audio data

Lng	WER (%)	MAP (iv)	MAP (oov)	MAP (rare)	MAP (targ)
Eng	35.96	0.846	0.621	0.866	
Rus.	62.34	0.614	0.325	0.638	0.412
Ukr.	56.43	0.702	0.631	0.715	0.404

At test time, we derive audio files from the provided videos. The videos are in .mp4 format, and we extract the audio track in .wav format using the free, open source tool

FFmpeg[1]. We convert the resulting .wav files to 16 KHz, 16-bit, NIST sphere format, the preferred format for BBN's speech recognizer.

For each audio file, we produce a transcript in which each word is annotated with its start time and duration. This represents our one-best output. We produce a consensus net containing many possible transcription outcomes, but store and pass only the one-best path for now.

4 Machine Translation

The BBN Neural Machine Translation (NMT) system employs a standard 6-layer Transformer model [1] jointly trained over Russian, Ukrainian, and English data. We used the *tensor2tensor* toolkit for the transformer implementation. We perform sub-word tokenization with the *sentence-piece* toolkit, an unsupervised text tokenizer that enables us to train an MT model without running language-specific tokenizers for Russian or Ukrainian. The sub-word vocabulary is shared between Russian, Ukrainian, and English and has a vocabulary size of 13,000. The primary data source is from the LORELEI program, augmented with a variety of web data such as CommonCrawl[2] and the open parallel corpus[3]. In addition, we add parallel sentences extracted from the headlines in the scenario document to the MT training data.

In total, we use 100 million Russian words and 9 million Ukrainian words in training. To correct the language imbalance, we duplicate the Ukrainian data so that the transformer model is exposed to an equal amount of Russian and Ukrainian data during training. This has been shown to be effective when building multi-lingual NMT models [2].

We built two variants of the MT system. We use a typical MT model as described above to process regular text data, and an ASR-variant to process textual data from ASR output. The BBN Speech-to-Text module produces output with no casing information, and minimal punctuation. To account for this, we trained a second MT model for which the input data was uppercased, and punctuation marks removed. We succeeded in producing case-variant and properly punctuated outcomes with this model; for this reason, we process Ukrainian, Russian, *and English* ASR output through our MT module. Language ID happens implicitly in the single, multi-lingual MT model.

MT training for each MT variant takes 8 hours on two Tesla V100 GPUs. BLEU scores on the LORELEI test sets are shown in Table 3.

Table 3: BLEU scores for BBN NMT

Language	BLEU
Ukrainian	20.4
Russian	33.1

We retain only sentence-based alignments of translated to original text. In the case of text transcribed from audio, we are able to correlate the translated sentence to its original time stamps (and thus video key frames). We were not able to implement word-based alignment over the results of NMT in time for the 2019 AIDA evaluation; for this reason, entity, relation, and event extractions from foreign-language and transcribed texts use the offsets of the sentence boundaries, rather than the specific word boundaries.

5 Text-Based Extraction

We use a variety of supervised models to extract named entities, relations, and events from English-language text (transcribed and/or translated from speech or foreign language, as appropriate).

For named entity recognition, we use SERIF, which applies a discriminative Viterbi-style perceptron model to find and extract names of

[1] https://ffmpeg.org/

[2] http://commoncrawl.org

[3] http://opus.nlpl.eu

persons, places, and organizations [3]. Mentions are grouped into entities using a sieve-based approach [4].

SERIF also extracts a set of relations with a maximum entropy model combined with heuristics. We supplement this with a second relation finding system that applies syntactic patterns expressed using the Brandy pattern language over pairs of entities as detected by SERIF. We authored a set of patterns specific to the AIDA scenario using examples from practice corpora. These cover a total of 30 different types in the AIDA ontology, including subtypes and sub-sub-types. Using the LearnIt tool, all pairs of entities from sentences inside the English documents were identified and used as potential examples for patterns. The LearnIt tool allows the user to query for text trigger words that identify potential relations.

We extract events using SERIF's logistic regression models as well as ACCENT, which identifies additional events and their matching arguments according to the CAMEO event ontology [5]. ACCENT finds events using structured patterns applied to augmented text graphs (normalized proposition trees that have been augmented with synonymy and coreference). The classes of events, relations, and entities extracted by SERIF and ACCENT are associated with the AIDA ontology via a manually-produced mapping between the ACE and CAMEO ontologies and the AIDA ontology.

A third approach to extracting events, a system we call NLPLingo, also leverages SERIF's named entity extractions, but uses a pair of convolutional neural networks (CNNs) to extract trigger words and associate likely arguments with those triggers to form events. Potential event trigger words in the text are selected according to their part of speech tags. The first CNN identifies any event types that can be associated with each trigger and assigns a confidence to each. The second CNN ingests these events and SERIF's entities and identifies the subset of nearby entities that fill particular role slots for each event.

For features, both networks use the word embeddings of a given chunk of the text, as well as those of the local context around candidate tokens for triggers and slot fillers. The trigger model uses the IOB-style entity type tags of each token in the sentence as a feature. The argument model also uses the predicted event type of the event trigger under inspection.

Event models are trained on all annotated data made available by LDC for the AIDA program, as well as LDC's Automatic Content Extraction (ACE) 2005 dataset. We used 90% of the available data for training, and reserved 10% for testing, resulting in respectable event extraction accuracy, as shown in Table 4. The gold-standard annotations have at most 1 type per trigger. For the AIDA program, we are interested in multiple hypotheses, and we turn to metrics from information retrieval to support this effort. In this case, Mean Average Precision up to *k* (MAP@*k*) is a measure of where the true type is located in the list of *k* type predictions ordered by decreasing confidence, averaged across triggers. MAP@1 is a measure of how often the predicted type with the highest confidence is correct, while MAP@T is a measure of how often the correct type is found among all predictions for a trigger. Average R-precision (ARP) is the proportion of the actual triggers for a given type which were assigned a higher confidence than all other triggers, averaged across types.

Table 4: Event extraction scores, held-out English AIDA data

Model	Micro F-1	MAP @1	MAP @T	ARP
Event type	0.67	0.679	0.694	0.448
Arguments	0.54	0.588	0.613	0.404

6 Image-Based Extraction

6.1 Facial Recognition

We apply the same facial recognition (Face ID) algorithm used in the previous AIDA evaluation. For the 2019 application, we expanded our gallery of known faces from 27 to 295, with help from a human annotator. Starting with names of people in the scenario document and practice annotation, we found all other names in a 1- or 2-hop relationship to the original names in DBpedia. Names collected from DBpedia were then restricted to ones that have a relation to Russia, Ukraine, Belarus, Donbass, or Luhansk. We also manually added some contemporary world leaders.

We use Face ID to detect persons of interest in images (png, jpg, bmp, gif) and videos (mp4). We adapt an open source implementation of FaceNet for this work [6].

Training occurs independently for the 3 stages of the FaceID pipeline. The three stages are: face detection, face image vector embedding, and face identification. The Multitask Cascaded Convolutional Network (MTCNN) face detection approach [7] was trained on the CelebA [8] and WIDER FACE [9] datasets. The learned model parameters for the MTCNN were obtained from the open source FaceNet implementation website[4]. Model parameters for the Inception Resnet v1 deep convolutional neural network architecture were trained on the VGGFace2 dataset [10].

The k-nearest neighbors within a Euclidean distance threshold are retrieved from the gallery using the FLANN library [11]. In Task 1a, majority voting based on these retrieved vectors from the gallery determine the identity of the query face vector.

A confidence score is computed between the query face vector and each one of its nearest neighbors. This confidence score is the result of applying a radial basis function kernel to the query vector and a vector in the gallery. The greatest sum of these confidence scores among the different gallery identities is used to select the identity of the query vector. The average of these confidence scores, per the selected identity, is used to report the confidence in the justification. If the average confidence is below a certain threshold, the system did not add a justification for that identity.

We are able to incorporate information from downstream hypotheses in Task 1b to alter the outcome of the Face ID analytic. For any person entity found in the hypotheses with a known KB link, if they are also found in our gallery of faces, we boost by a constant the probability of that name as an outcome across all detections. This results in a few more detections of those names than were seen before the hypothesis-based boosting.

6.2 Concept Detection

We train a set of video concept detection models from open source data. Video concepts may refer to contexts, objects, or situations. These are mapped to the AIDA ontology (manually) as events, relations, or entities.

We train multiple concept detectors using deep convolutional network models that have been fine-tuned to detect scenario-relevant concepts. For training, we use a subset of the OpenImages[5] dataset that includes 111 concepts relevant to the AIDA scenario.

We train a multi-label convolutional network (MLCN) for scenario-relevant concepts using the CAFFE toolbox[6]. The MLCN has a similar network structure as in the fully convolutional VGG16 network in [12] and includes two includes two parts: a CNN and a Multi-label Classifier. The CNN includes 15 convolution layers and 5 max-pooling layers, while the

[4] https://github.com/davidsandberg/facenet

[5] https://storage.googleapis.com/openimages/web/index.html

[6] http://caffe.berkeleyvision.org/

Multi-label Classifier is modeled as a convolution layer with K kernels of size 1x1 and a sigmoid layer. We also add the multiple instance learning (MIL) layer provided by [13] to the MLCN. The MIL layer pools together the CNN features computed on the image regions spatially. We trained the MLCN for 111 concepts with 27,198 images, where the MLCN is initialized by ImageNet pre-trained weights. We evaluate the model on a validation set which includes 4,296 images, with a result of 62.6% mean average precision.

In addition to our robust MLCN model, we use a pre-trained object detection model of Atomic Visual Actions (80 actions) and a second pre-trained object detection model trained on a subset of OpenImages classes utilizing the Tensorflow toolbox. We map the concept detections to entities and other ontology elements, with bounding box and associated key-frame information. Our system currently maps detections from these model outputs to the AIDA ontology using predefined mappings.

We have investigated deep learning models that jointly detect objects using image-caption pairs. Using a pre-trained semantic embedding network that uses a CNN for image embedding and a RNN for embedding a sentence/text snippet, we learn a common embedding for image-caption pairs and use the embedding to classify the objects in images. This has resulted in multiclass classification improvements over image-only models in a small annotated dataset from AIDA seedling corpora; for instance, the Micro Average improves by 0.0643 absolute F1 score. However, the overall scores remain low, due to a mismatch between the training set and the small AIDA test set. We expect that further curation of the data sets will result in better overall classification, as well as continued improvements from including the text embedding.

In addition, we are in the process of adapting landmark detection models to create a Landmark ID module for identifying well-known landmarks which will allow us to extract location information from image and videos.

7 Cross-modality merging

The extractions provided by each of the text-based and image-based analytics described above are converted to the required AIDA Interchange Format (AIF) to include name strings and appropriate justifications. We perform cross-modal co-reference for each parent document by matching name strings of entities and event types. We also perform entity linking to the AIDA knowledge base, providing a consistent ID string for any entity found in that resource. The combined information for an entity or an event is called a Knowledge Element (KE) and includes justifications from images, videos, and text (including transcribed text). We choose an *informative mention* for each knowledge element by preferring canonical mentions produced by SERIF for any text justifications present for the KE. If no such justification exists, then we choose the mention with highest confidence across all remaining analytics.

8 Conclusion

For the SM-KBP track of TAC 2019, BBN produced a set of knowledge graphs consisting of elements drawn from text, video, and audio sources using a variety of analytic components trained on open-source and curated scenario-relevant resources. We combine neural and knowledge-rich approaches for event and entity extraction. A push-button process, which also includes AIF validation and production of a simple HTML display of results for each parent document, has been implemented.

9 Acknowledgements

This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA). The views, opinions and/or findings expressed are those of the author and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government. Distribution ‘A’: Approved for Public Release, Distribution Unlimited.

10 Bibliography

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is All You Need," in *NIPS 2017*, Long Beach, CA, 2017.

[2] M. Johnson, M. Schuster, Q. V. Le, M. Krikun, Y. Wu, Z. Chen, N. Thorat, F. B. Viégas, M. Wattenberg, G. Corrado, M. Hughes and J. Dean, "Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation," *TACL*, vol. 5, pp. 339-351, 2017.

[3] L. Ramshaw, E. Boschee, M. Freedman, J. MacBride, R. Weischedel and A. Zamanian, "SERIF Language Processing - Effective Trainable Language Understanding," in *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*, Springer, 2011, pp. 626-631.

[4] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu and D. Jurafsky, "Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task," in *Proc. 15th Conf. on Computational Natural Langauge Learning: Shared task*, 2011.

[5] E. Boschee, P. Natarajan and R. Weischedel, "Automatic Extraction of Events from OpenSource Text for Predictive Forecasting," in *Handbok of Computational Approaches to Counterterrorism*, 2012, pp. 51-67.

[6] F. Schroff, D. Kalenichenko and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," in *arXiv:1503.03832*, 2015.

[7] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Neural Networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499-1503, 2016.

[8] Z. Liu, P. Luo, X. Wang and X. Tang, "Deep Learning Face Attributes in the Wild," in *Proc. Int'l Conf. on Computer Vision (ICCV)*, 2015.

[9] S. Yang, P. Luo, C. C. Loy and X. Tang, "WIDER FACE: A Face Detection Benchmark," in *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[10] Q. Cao, L. Shen, W. Xie, O. M. Parkhi and A. Zisserman, "VGGFace2: A dataset for recognising face across pose and age," in *Int'l Conf on Automatic Face and Gesture Recognition*, 2018.

[11] M. Muja and D. Lowe, "Scalable Nearest Neighbor Algorithms for High Dimensional Data," *Pattern Analysis and Machine Intelligence*, vol. 36, 2014.

[12] E. Shelhamer, J. Long and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640--651, 2015.

[13] H. Fang, S. Gupta, F. N. Iandola, R. K. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick and G. Zweig, "From captions to visual concepts and back," in *CVPR*, 2016.