Updates and Analysis of BBN Panorama for SM-KBP

Roger Bock, Jordan Hashemi, Ilana Heintz, and Ben Rozonoyer
Raytheon BBN Technologies
Cambridge, MA 02138, USA

Abstract

We provide system updates and performance analysis regarding the 2020 version of the BBN Panorama multi-modal processing pipeline, as submitted to the 2020 Streaming Media Knowledge Base Population track.

1 Introduction

BBN developed and deployed the Panorama multi-modal pipeline for participation in the 2020 TAC SM-KBP challenge. The workflow is depicted in Figure 1: a "parent" document is recorded in the document database, with metadata about each of the "child" documents. In the 2020 version, we process only text and images, as resources did not allow for annotation of speech data, resulting in false positives for any speech-based extractions. BBN neural machine translation, a multi-lingual model, processes all non-English data to produce an English outcome.

These translated documents and all English text are first passed to *SERIF*, BBN's natural language processing stack, which uses a variety of supervised algorithms to perform named entity recognition, relation extraction, and event and event argument extraction. SERIF annotations are the basis of additional processing by: *ACCENT*, which applies a set of propositional patterns to find events associated with the CAMEO ontology; *NLPLingo*, a CNN-based event extraction and event argument attachment engine, and *FactFinder*, a pattern-based approach to relation and event extraction.

Separately, image data is passed through a person recognition component that leverages the FaceNet implementation of a multi-task

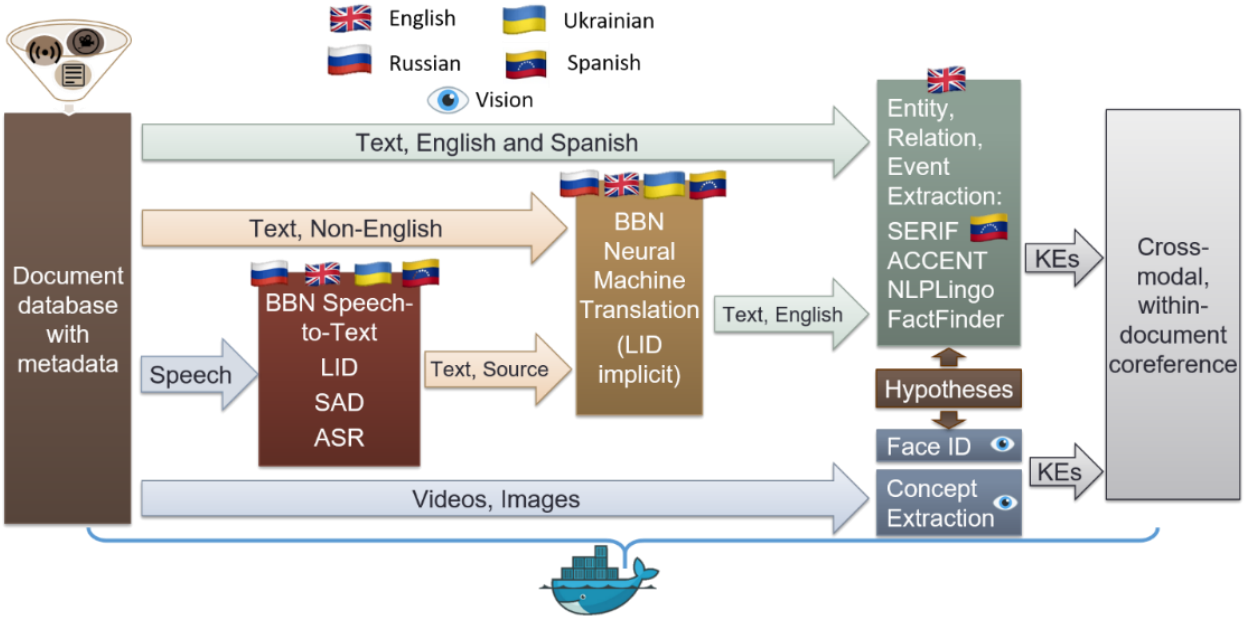

Figure 1: The BBN Panorama pipeline processed text and visual images to produce knowledge elements for the 2020 SM-KBP track.

cascaded CNN (MTCNN) [1], as trained on the CelebA [2] and WIDER FACE [3] datasets, and Inception Resnet deep CNN architecture as trained on the VGGFace2 dataset [4]. We created a gallery of 50 persons relevant to the Venezuela scenario.

A "merger" step uses heuristics to create co-reference chains among the entities found by text NER and image-based face recognition, and across the relations and events found by the various text extraction modules, across all child documents associated with a single parent. The outcome is a set of "knowledge elements" in the AIDA Interchange Format for each parent document.

2 System Enhancements

We describe changes made to the Panorama system in 2020 to accommodate new concepts and languages, and provide more informative knowledge elements to downstream processes.

2.1 Automatic Speech Recognition (ASR)

Although this year's TAC challenge did not include assessment of speech data, we were prepared to process it with a Spanish-language transcription model. We trained the model using a time-delay neural network with multilingual initialization similar to the architecture described in [5]. We then evaluated the word error rate (WER) of our models for each of the languages on a general domain test, where lower WER is better. As shown in Table 1, English performs the best, which is expected, given the availability of significantly more training data.

Language	# Hours in Training Data	WER (%)
Russian	50	31.4
Spanish	54	35.8
English	2300	12.1

Table 1 - Baseline General Domain ASR Performance

Using the information available in the *Crisis in Venezuela* scenario document, we scraped scenario-relevant text data from Wikipedia to augment the language models used by our ASR system. We identified and scraped the contents of relevant English Wikipedia pages. Where there existed corresponding Russian and Spanish pages, we also scraped those.

We repeated the experiments from Table 1 using the augmented language models. In Table 2, we report WER on the general domain test sets for each language, and compare performance with and without the inclusion of the additional language modeling data from Wikipedia. We see a small degradation in WER on the general domain test sets when using the augmented language models, which we expect is due to the domain mismatch.

Language	# Hours in Training Data	Wikipedia Data	WER (%)
Russian	50	N/A	31.4
Russian	50	31 articles	32.0
Spanish	54	N/A	35.8
Spanish	54	33 articles	36.6
English	2300	N/A	12.1
English	2300	35 articles	13.4

Table 2 - General Domain ASR Performance Using Wikipedia-Augmented Language Models

2.1.1 Machine Translation (MT)

Following a similar approach to modeling for the Ukraine-Russia Relations scenario, we trained a single multi-way machine translation model for the Crisis in Venezuela scenario. Specifically, we trained a model that can take as input mono-cased English/Russian/Spanish or true-cased Russian/Spanish and produce true-cased English as output. The inclusion of mono-cased data as input allows us to handle the mono-cased text produced by ASR.

The model architecture we are using is an encoder-decoder transformer with self-attention; similar to what is described in [6]. We evaluated the model's performance on datasets from the LORELEI program. Table 3 shows the BLEU scores on these test sets. Spanish performance appears significantly better than Russian, and in general the BLEU scores are satisfactory for use in the Panorama pipeline.

Corpus	Lang	BLEU
LORELEI_RUS_Parallel_Found_test	Rus	34.5
LORELEI_RUS_Parallel_From_RUS_test	Rus	33.0
test.lorelei.spanish.from_spa.v1	Spa	43.0

Table 3 - BLEU scores on in-house test sets developed under the LORELEI program

2.1.2 Entity Linking – M36 Evaluation

In the previous version of Panorama, we provided single-best entity links for people found by our Face ID system and for textual entities that mapped to GeoNames. An improved version will include an entity linking model that returns a distribution over entity links for all entities found in text.

We are basing our approach on that of [7], where many "word expert" models are trained instead of a single monolithic model. Mentions are represented as the average of the embeddings of neighboring words in a context window. The classification model consists of two fully connected layers followed by a softmax layer. We train one of these models for every possible ambiguous mention.

We use a 2014 snapshot of Wikipedia to train our models, where the hyperlink anchor text defines the set of possible candidate mentions, and the hyperlink context and the linked page serve as the training data. We limit ourselves to the 523K mentions strings that occur more than ten times as anchors, so we learn 523K separate classifiers.

Traditional approaches to entity linking learn a single model that does N-way classification over all N entities in a knowledge base. The disadvantage of these monolithic models is that they can take tens of days to train on high end GPU machines. By learning one word expert model for each of M mentions, we end up with many small models, which can be trained in parallel on commodity CPU machines. This results in faster training and means that adaptation to new domains can be

"Republican presidential candidate John McCain won the Washington primary on Tuesday"

Score	Enitity
98.2%	Washington_(state)
1.3%	Washington,_D.C.
0.006%	University_of_Washington
…	…
0.0000007%	Washington_County,_Ohio

"The ADP had already grabbed all 13 seats put to the ballot in a first-round of the elections"

Score	Enitity
24.9%	Anguilla_Democratic_Party
20.9%	Arab_Democratic_Party_(Lebanon)
14.8%	Arab_Democratic_Party_(Israel)
…	…

Figure 2 - Distributions over entity links for two example mentions

done quickly, as only affected models need to be retrained.

Figure 2 shows sample output from our entity linking system for two mentions. For the mention "Washington" in the first sentence, the model produces a very peaked score distribution for the entity corresponding to Washington State. For the mention "ADP" in the second sentence, the model produces a much flatter distribution, presumably because the context does not provide as much evidence to disambiguate the entity.

2.1.3 Entity Linking – Post-Evaluation

To support increased recall and precision in downstream hypothesis generation modules, we updated our system after the formal evaluation to include links from entity mentions to unique identifiers in the WikiData repository.

We considered different published approaches, including: training separate classifiers for each mention string [7]; using BERT to encode mentions and entities in the same space for zero shot linking [8]; treating entity linking as a sequence labeling task with BERT [9]; and using a bi-encoder for candidate generation and a cross-encoder for reranking [10].

We implemented the approach taken by [10] due to the state-of-the-art performance on TACKBP-2010 and simultaneous support for zero-shot entity linking. The latter capability will be useful for linking against the scenario-specific entities provided in the augmented KB.

Briefly, our model learns to project mentions in context and the entities they refer to into the same area of a high dimensional space. Prior to projection, it represents mentions in context as a sequence of tokens, with the mention surrounded by context on either side and demarcated by special tokens. It represents entities as a concatenation of the entity's name and a brief description. After mention projection and identification of neighboring candidate entities, a more expensive reranking step uses a transformer model that can attend to both the mention representation and the entity representation to score the candidate entities.

We apply this model to every TextJustification produced by Panorama, and report entity links where the linking score exceeds a threshold of zero (scores can be negative), which strikes a reasonable balance for reporting a useful distribution of entity links. Table 4 shows some sample output from our WikiData linking system. Note that in addition to the traditional linking of named entities, we also link event triggers (e.g., manifestar) to their corresponding pages.

Mention	Context	Predicted Wikidata QIDS and scores
San Crístobal	... collectives were patrolling the streets of the state capital San Crístobal.	Q820235 [San Cristóbal, Táchira]: 6.07 Q2647967 [San Cristóbal, Bogotá]: 0.70 Q2884933 [San Cristóbal de las Casas]: 0.35
National Assembly	... in particular the president of the opposition-led National Assembly, Julio Borges.	Q1585014 [National Assembly (Venezuela)]: 6.67 Q1969591 [National Assembly (Nicaragua)]: 1.79 Q1319595 [National Assembly (Ecuador)]: 0.99
manifestar	... sigue midiendo la disposición de las personas a manifestar.	Q175331 [Demonstration (political)]: 7.02 Q273120 [Protest]: 6.96 Q1395149 [Demonstration (teaching)]: 1.79

Table 4: Mentions in context and their corresponding predicted WikiData QIDs

2.1.4 Knowledge Element Embeddings

We calculate Knowledge Element (KE) embeddings using BERT, similar to face embeddings from Face ID and event embeddings from NLPLingo, in an effort to communicate more robust contextual information to downstream modules.

After calculating contextualized BERT embeddings for each token in a text document (English or machine translated), we compute an embedding for each KE by averaging the BERT embeddings of all of the tokens comprising its most informative justification. We expect this information from the BERT-based vector representations of KEs to aid in corpus-level entity and event coreference resolution.

3 Extracted Type Coverage and Analysis

We used the dry run data from NIST and LDC to perform a close analysis of the extraction accuracy across types. At the sub-subtype level, the AIDA ontology has 179 entity types, 50 relation types, and 149 event types. We extract these types using a combination of new and existing models, including feature-based and rule-based algorithms for all types, and deep neural networks for event and event arguments. The models do not cover all of the ontology sub-subtypes. Within the annotation set, we determine the following type coverage:

- 99.6% of entity mentions
- 98.6% of relation mentions
- 93.7% of event mentions

These figures represent the upper bound of our ERE recall in this data subset.

Looking more closely at event extraction, a key aspect of the AIDA program, we note that our event models identify arguments irrespective of the type restrictions in the AIDA ontology. For instance, from the sentence,

> *"Whoever the attackers were, Mr Rodriguez said they had "failed" – although seven of the National Guard were injured and have been receiving treatment."*

Our system extracts a *Medical.Intervention.Intervention* event with a *Patient* argument covering the span *"seven of the National Guard".* In the AIDA ontology, Patients are required to be PER (person entities) and our system typed this mention as an ORG. As a result, this argument is not produced as part of the outcome. We observed similar behavior in other event types, such as *Conflict.Coup.Coup* where the deposed entity can be an ORG, but not a GPE, which is how the system often (reasonably) tags it. To improve results within the current AIDA ontology, we will consider augmenting the entity with an additional type that permits the event argument attachment; further testing would illuminate how the higher recall compares to a probable increase in false positives.

Place arguments are an important aspect of events for hypothesis generation. Panorama attaches a Place argument to 35% of the extracted events of the AIDA 2020 dry run corpus (LDC2020E11). Many of these are not likely to serve as entry points to the hypotheses, for instance, "house" and "balcony." A key place descriptor such as a city name often appears near the event trigger, but not in the same sentence, and is not attached to the event as a Place argument. To improve Place attachment for better near-term results in hypothesis generation, we implement a heuristic approach. In previous projects, we have used a set of heuristics to identify whether a document is predominately about a single country, and if so, to identify that country. This has been successfully applied for disambiguation of other mention types, such as "Labour Party". We find that 88% of the documents in the data subset include a single country that we can identify as primary using

these heuristics. The same heuristics may be applied to cities or states, as well. Provision of this information as a Place argument for events should increase recall and depth of analysis for hypotheses.

Event times are equally important to hypothesis generation. We note that the Panorama pipeline attaches dates to only 9% of the events found in the dry run corpus. Our intuition and anecdotal evidence from the corpus tell us that most documents discuss events that have happened in the very recent past. As a stopgap measure, to provide useful, if not highly precise, information to downstream partners, Panorama will provide a date with each event that refers to a window of time a few days previous to the publication date.

We incorporated an extensive list of drones (*VEH.Aircraft.Drone* in the AIDA ontology) from Wikipedia into a regular-expression-based entity typing module. Most scenario-specific documents that mention drones will more frequently mention the word “drone” than the specific make of the drone (provided that it is not a common confounding word such as “bird” – we have excluded ambiguous names from the list). For the sake of recall, our recognizer captures phrases containing the word “drone” and its inflections; we expect that the verb sense of “drone” will not occur frequently enough in scenario-related documents to hurt precision.

4 Conclusion

We continue to improve the Panorama multi-modal knowledge element extraction tool to supply knowledge graph and hypothesis generation algorithms the necessary data to drive their research.

5 Acknowledgements

This research was developed with funding from the Air Force Research Lab (AFRL) and 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. This report has been Approved for Public Release, Distribution Unlimited.

6 Bibliography

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] F. Keith, W. Hartmann, M.-h. Siu, J. Ma and O. Kimball, "Optimizing multilingual knowledge transfer for time-delay neural networks with low-rank factorization," in *ICASSP*, 2018.

[6] F. Wu, A. Fan, A. Baevski, Y. N. Dauphin and M. Auli, "Pay less attention with lightweight and dynamic convolutions," arXiv preprint , arXiv: 1901.10430, 2019.

[7] A. Barrena, A. Soroa and E. Agirre, "Learning text representations for 500K classification tasks on Named Entity Disambiguation," in *Proc. 22nd Conf on Computational Natural Language Learning*, 2018.

[8] 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.

[9] 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.

[10] 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.

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

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.