A Deep Learning Approach for CAMEO Coding

LUCAS
August, 2017

Abstract

We proposed a deep learning approach for coding event data with CAMEO codes. Current coding engines including TABARI and PETRARCH require compilation of specific dictionaries while our deep learning models only requires pre-labelled data. We showed that our deep learning models outperform PETRARCH and other machine learning models and the hierarchical network is the most suitable tool for CAMEO coding both because of its accuracy and practical meaning of the classification. Our deep learning models can also be easily adapted to an expanded version of CAMEO codes and can be transplanted to improve the current event coding engines and other projects.

I. Introduction

Political event data is useful in quantitative political analysis. To collect, categorize and utilize event data extracted from news articles, a coding standard is necessary. Several ontologies have been developed for coding event data, including COPDAB by Azar (1980), WEIS by McClelland (1976) and the most recent one, CAMEO (Conflict and Mediation Event Observations) by Schrodt[1] and Yilmaz (2002). In its original version, there are 20 base categories and 296 subcategories in total, which makes manually labelling almost impossible. According to Schrodt and Brackle (2013), humans are capable of coding about six to ten events per hour. The low efficiency and high cost of human labor necessitate an automated coding system. Efforts in this direction have been made since 1990s. A detailed review can be found in Schrodt, Beieler and Idris (2014). There are two popular automated coding systems for CAMEO code: the NSF-funded TABARI and its successor, python-based PETRARCH. Compared with human labor, these automated coding systems demonstrate unprecedented efficiency in coding event data, at a speed of almost 2000 per second (Schrodt and Brackle, 2013).

One of the major task in coding event data is to classify the given sentence to correct category. This problem belongs to the realm of text classification. Machine learning has been widely applied to this field in the past twenty years, while deep learning has become prevalent only in recent years. These efforts include the application of hierarchical neural networks (HNN) (Ruiz and Shrinivasan, 1999, 2002), convolutional neural network (CNN) (Kim, 2014), Recurrent neural network (RNN) (Lai, Xu, Liu, and Zhao. 2015), or some other combination of those two models. To its nature, CAMEO coding is a difficult problem due to the large number of CAMEO codes, or labels. This kind of problem is addressed as extreme multi-label text classification (XMTC) and remains as a challenge. Liu *et al.* (2017) reviewed the recent progress in XMTC. Another difficulty in CAMEO coding results from the ambiguous meaning of the sentence and

[1] We have consulted the history of automated coding systems with Professor Schrodt and we thank him for his detailed explanation and his contribution in this field.

the labels. Many CAMEO codes have similar meaning which is hard to distinguish even by humans. For example, CAMEO code 0211 represents the action 'Appeal for economic cooperation' while 1011 represents the action 'Demand economic cooperation'. These challenges in CAMEO coding necessitate the needs for applying an automated, deep learning based approach to this task. To the best of our knowledge, we are the first to work in this direction.

Current automated coding systems, including TABARI and PETRARCH, require the compilation of detailed dictionaries of actors, agents, verbs and other necessary information to recognize and categorize event data. Professor Schrodt and his team have devoted significant effort to compiling those dictionaries, which are open-source and publicly available. This method of categorizing sentences requires not only a large investment of human-labor hours, but also subject matter expertise. In their application, case-specific dictionaries of actors and agents in certain areas were developed to improve model accuracy in identifying event data for a particular affair. Therefore, it is difficult to extend the engine for a specific use in other regions or adapt it to an expanded coding system. In contrast, a deep learning approach does not require compiling dictionaries, saving many human labor hours and cost. Meanwhile, it is easy to expand the coding system by simply collecting corresponding sample sentences and training the neural network on the new data set. In our Hermes Project, we revise the original version of CAMEO codes[2] with 296 categories to an expanded version of 374 categories without spending time on finding key actors, agents and verbs of those new categories.

Thus, the deep learning approach for CAMEO sentence classification relieves analysts from extensive data entry, increases accuracy, and contributes enormous scalability to handle future expansion even beyond the domain of event data. Further, the efficiency of a neural network permits an analyst to code events scraped from the Internet, drawn from a text archive, or even handwritten on-command. Such a tool is intensely applicable to political science research and is a great boon to both broad-spectrum and case-specific data collection efforts.

II. Data

A deep learning approach for CAMEO coding requires pre-labelled sentences for training and testing. Collecting such sentences is very time and money consuming. However, there is no such publicly available dataset hitherto. The Phoenix Data Projects by Open Event Data Alliance and The Computational Event Data Project led by Professor Schrodt provide access to event data without original texts, hence those data cannot be used for training a machine learning algorithm. To solve the problem, we have manually collected and labelled 4,532 sentences for our expanded version of CAMEO codes with 374 categories. Meanwhile, to make it comparable with the original version, we have 3,587 sample sentences for the original 296 CAMEO codes. It should be noted that the sentence collection would more or less be affected by individual bias. A sentence can be categorized to different categories by different people. Nevertheless, these are the only data available to us and we will train and test our models and compare the performance with current coding systems.

---

[2] We refer to http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf as the latest version of standard CAMEO codes.

In our Hermes Project, we train our model on the expanded version of CAMEO codes with 374 categories; but the PETRARCH engine we want to compare with was built on the original version with 296 categories. Thus, all the models and test results in this paper are based on the original version. The samples sentences are not distributed evenly among all 296 categories. On average, we have 12 sentences for each category while the maximum is 213 (CAMEO code 010: 'Make a statement') and the minimum is 2. The scarcity of data limits the performance of our model. It will be gradually improved as we are collecting and labelling more sentences using our PipeLine Project. The PipeLine projects applies the open-source scraper from The Phoenix Project and uses our pre-trained neural network to give predictions for scraped sentences, then filters these predictions with human intervention in a graphic user interface.

III. Methods

We test different machine learning and deep learning models for the multi-label classification task of CAMEO sentences. A fully-connected neural network with one hidden layer is built as the baseline model. Due to the nature of CAMEO codes whose 296 categories are divided into 20 classes, we also construct a hierarchical neural network with one gating network that has 20 exits and 20 expert networks following the gating network. Other machine learning techniques including logistic regression, tree-based models, Naïve Bayes classifiers, KNN and SVM are also tested for this task as comparison. We have also run the PETRARCH engine on our test data set to compare the results. It turns out that the deep learning approach, especially the hierarchical neural network, outperforms other techniques for CAMEO coding. Thus, a deep learning approach for CAMEO coding can be applied to improve the engine for recording event data. In our experiments, all neural network models are built using Google's TensorFlow.

1. Fully-Connected Neural Network

The fully-connected neural network uses bag-of-word vector representation of sentence as its input. The basket of words is chosen from all the nouns, verbs and adjectives extracted from all 4,532 sample sentences; all words are in lemmatized form and lower case. We apply the Stanford CoreNLP engine in completing this task. Among all extracted words from sample sentences, we use the most frequent 3000 words as features to represent a sentence by a $3{,}000 \times 1$ vector. It is possible to choose the basket of another scale, but in our experiments 3,000 is the number that leads to the second highest performance while it is time-and-storage efficient. Figure 1 shows the top-5 accuracy using bag-of-words vectorization with different number of frequent words:

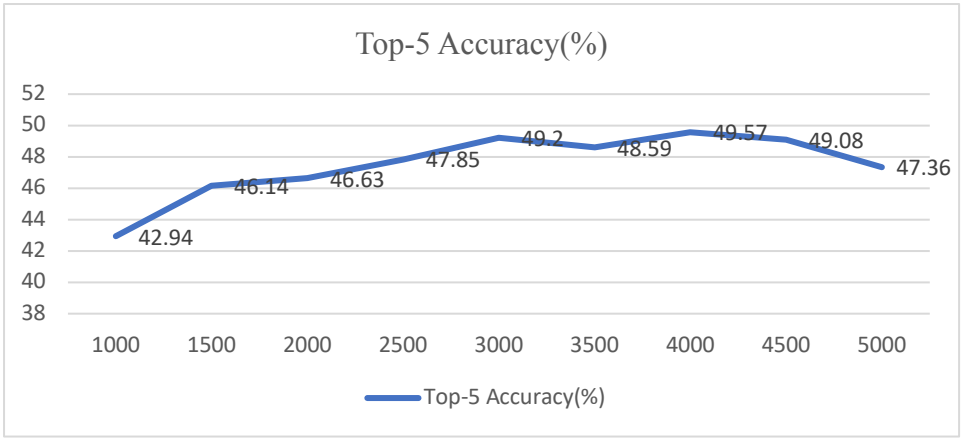

Figure 1 Top-5 accuracy using different number of bag-of-word representation

Our fully-connected network has one input layer with 3,000 nodes, one hidden layer with 500 nodes and one output layer with 296 nodes. Before the final output layer is a softmax activation layer:

$$\hat{p}_{ij} = \frac{\exp(\hat{y}_j(x_i))}{\sum_{k=1}^{L} \exp(\hat{y}_k(x_i))},$$

where $\hat{p}_{ij}$ is the predicted probability of sentence vector x_i on CAMEO code j, $\hat{y}_j(x_i)$ is original model prediction of x_i. As in Liu *et al.* (2017), we apply the classic cross-entropy loss function for multi-label classification:

$$\min_{\Theta} -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{L} y_{ij} \log(\hat{p}_{ij}),$$

in which Θ denotes parameters to be optimized. Other optimizations include using an exponential moving-average model and L2 regularization.

2. Hierarchical Neural Network

Hierarchical neural network is a suitable tool for the text classification, especially in multi-label classification task with scarce training data. A hierarchical neural network is actually not a single neural network but a cluster of neural networks organized in a tree-form. In a multi-label classification task, a hierarchical neural network will not categorize the instance directly to the final label, but instead will first classify the instance to a broad class through a gating network and use corresponding expert network for further classification. When labelled data are scarce, the gating network can take the advantage of reduced categories and increased sample size for each broad class in the first-level classification to improve the accuracy while allow detailed treatment under each broad class. A hierarchical neural network can have multiple levels, but in our experiment, we only use one parent level for gating network and one child level for 20 expert networks. The performance of hierarchical neural network undoubtedly depends on the division

of broad classes. In CAMEO coding, however, there is a natural division as the CAMEO codes are of 20 broad classes in the code book. This original division is based on the key verb rather than the content of a sentence.

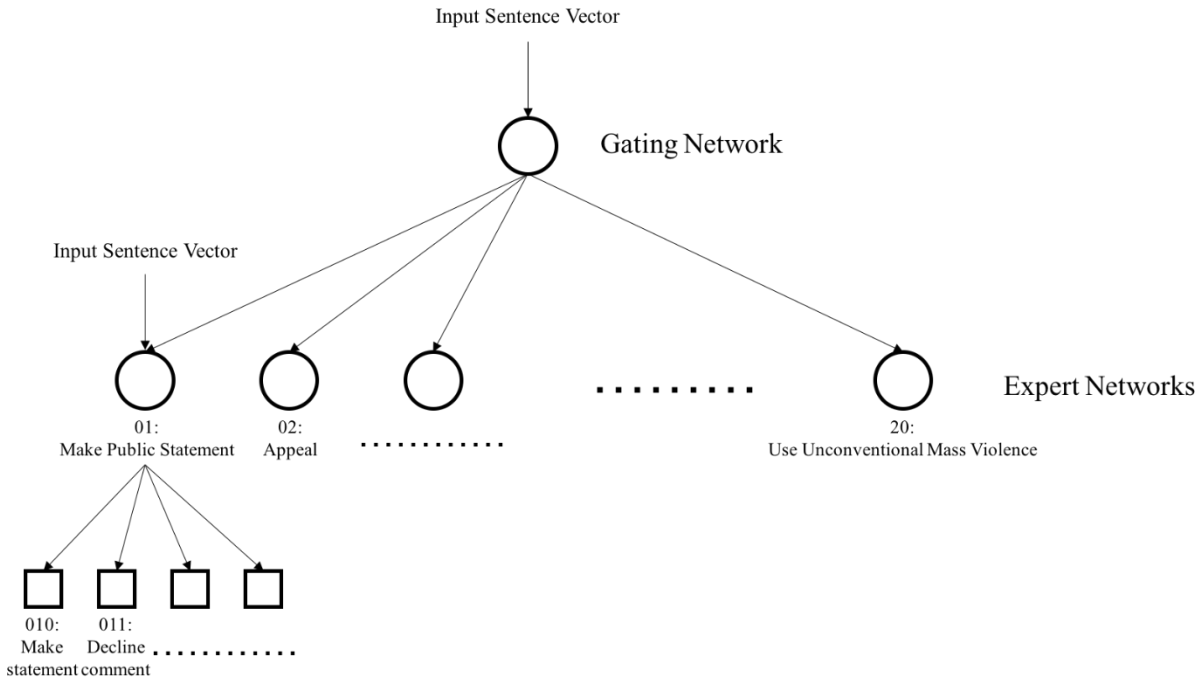

Figure 2 Hierarchical Neural Network

It is also possible to set different structure for each expert network. In our project, in order to allow an automated training of all neural networks, we set the same structure for every expert network which is a fully connected neural network similar to our baseline model but with 100 nodes in its hidden layer. By doing so, political scientists need only to focus on the division of CAMEO codes rather than the technical details.

To predict the top-n possible labels, the hierarchical model differs slightly with the fully-connected model, because there are different ways to get the final n predictions: we can either use the top-n predictions by expert network of the most possible broad class or use a combination of top predictions from several expert networks of top-m most likely labels predicted by the gating network. In our experiment, since we want to calculate the top-5 accuracy, we use the top-2 predictions by the gating network and adopts the first three predictions from expert network of the most possible label and the first two predictions from the expert network of the second most possible label.

IV. Experiments

We train our fully connected model and hierarchical model using 2684 records and test these models using 815 records. As comparison, we also train and test other machine learning models on the same data set. These experiments are presented in Table 1. From Table1, we can see that

the hierarchical neural network yields the highest absolute accuracy and Top 5 accuracy, which confirms that the hierarchical neural network is an ideal tool for CAMEO coding. Compared with our baseline model, the fully connected neural network, the hierarchical model has significantly higher absolute accuracy, while its top-5 accuracy is still a little bit higher than the baseline model. Meanwhile, the classification result by the gating network is also useful since it categorizes a sentence to the first level of CAMEOs codes. Since we use the first two predictions by the expert network, the top-2 accuracy of the gating network is 50%.

Table 1: CAMEO Classification Accuracy of Different Approaches

Model	Absolute accuracy	Top-5 accuracy
Hierarchical NN	33%	49%
Fully-Connected NN	24%	49%
Extremely Randomized Trees	24%	45%
Random Forest	23%	45%
Decision Tree	15%	17%
Gradient boosting tree	15%	22%
Logistic Regression	25%	49%
Bernoulli Naïve Bayes	5%	9%
Gaussian Naïve Bayes	10%	13%
K-nearest Neighbors	5%	10%
Support vector machine	5%	None*

* SVM does not support predicting probabilities of different classes

To our surprise, among traditional machine learning techniques, logistic regression has the highest performance whose top-5 accuracy is almost as high as the hierarchical model. In those tree-based models, extremely randomized trees have the highest performance. The Naïve Bayes models seem not to fit this task well.

Besides these machine learning models, we also compared our deep learning approach with the PETRARCH engine. However, according its design, the PETRARCH engine will discard a bunch of sentences that it conceives as containing no event data. We convert our 815 test sentences to XML file that can be read by the PETRARCH and let the engine classify these sentences. In our experiment, it recognizes 316 sentences and among these 316 sentences it gets 118 correct predictions, which is a 37.34% accuracy. Meanwhile, our baseline fully connected neural network gives 127 correct predictions, which is a 40.19% accuracy. Therefore, even on the narrowed data set preferred by PETRARCH, our deep learning model outperforms the PETRARCH.

V. Conclusions

Deep learning models are proper tools for CAMEO coding task. To its nature, the hierarchical neural network that divides the original version of 296 CAMEO codes to 20 broad class based on the code book yields the highest performance in all machine learning models we test. A deep learning model for CAMEO coding can be easily expanded to adapt new CAMEO codes, while

the current PETRARCH requires the compilation of corresponding special dictionaries. On sentences that can be recognized by PETRARCH, the deep learning model also has higher performance. Meanwhile, deep learning models are capable of transfer learning and can be transplanted to other projects. Since PETRARCH is primarily for recording event data, it will discard a large number of sentences, while our deep learning models are designed for giving predictions for arbitrarily given sentence. It is also possible for our models to discard records, so long as we add a label as 'discard' and provide discarded sentences as training data.

The performance of these deep learning models will be increasing as training samples are scaled up. In our initial model for he expanded version of 374 CAMEO codes, the absolute accuracy increases from 18% to 22% and the top-5 accuracy increases from 38% to 45% with 800 additional training samples. Due to limited data, convolutional neural network that requires a vector representation of word performs badly in our experiment. As we are continuing collecting more data, the CNN for CAMEO coding with architecture similar to Kim (2014) will be developed.

References

Gerner, Deborah J., et al. "The creation of CAMEO (Conflict and Mediation Event Observations): An event data framework for a post cold war world." *annual meeting of the American Political Science Association*. Vol. 29. 2002.

Gerner, Deborah J., et al. "Conflict and mediation event observations (CAMEO): A new event data framework for the analysis of foreign policy interactions." *International Studies Association, New Orleans* (2002).

Hughes, Mark, et al. "Medical Text Classification using Convolutional Neural Networks." *arXiv preprint arXiv:1704.06841* (2017).

Kim, Yoon. "Convolutional neural networks for sentence classification." *arXiv preprint arXiv:1408.5882* (2014).

Liu, Jingzhou, et al. "Deep Learning for Extreme Multi-label Text Classification." *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2017.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55-60.

Ruiz, Miguel E., and Padmini Srinivasan. "Hierarchical neural networks for text categorization." *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999.

Ruiz, Miguel E., and Padmini Srinivasan. "Hierarchical text categorization using neural networks." *Information Retrieval* 5.1 (2002): 87-118.

Schrodt, Philip A., and David Van Brackle. "Automated coding of political event data." *Handbook of computational approaches to counterterrorism*. Springer New York, 2013. 23-49.

Schrodt, Philip A., and Deborah J. Gerner. "Validity assessment of a machine-coded event data set for the Middle East, 1982-92." *American Journal of Political Science* (1994): 825-854.

Schrodt, Philip A., John Beieler, and Muhammed Idris. "Three'sa Charm?: Open Event Data Coding with EL: DIABLO, PETRARCH, and the Open Event Data Alliance." *ISA Annual Convention*. 2014.

Wang, Peng, et al. "Semantic Clustering and Convolutional Neural Network for Short Text Categorization." *ACL (2)*. 2015.

Yang, Zichao, et al. "Hierarchical Attention Networks for Document Classification." *HLT-NAACL*. 2016.