

Detecting Potential Topics In News Using BERT, CRF and Wikipedia

Swapnil Ashok Jadhav, Dailyhunt

Abstract

For a news content distribution platform like Dailyhunt¹, Named Entity Recognition is a pivotal task for building better user recommendation and notification algorithms. Apart from identifying names, locations, organisations from the news for 13+ Indian languages and use them in algorithms, we also need to identify n-grams which do not necessarily fit in the definition of Named-Entity, yet they are important. For example, "*me too movement*", "*beef ban*", "*alwar mob lynching*". In this exercise, given an English language text, we are trying to detect case-less n-grams which convey important information and can be used as topics and/or hashtags for a news. Model is built using Wikipedia titles data, private English news corpus and BERT-Multilingual(Devlin et al., 2018) pre-trained model, Bi-GRU(Chung et al., 2014) and CRF architecture. It shows promising results when compared with industry best Flair², Spacy³ and Stanford-caseless-NER⁴ in terms of F1 and especially Recall.

1. Introduction & Related Work

Named-Entity-Recognition(NER) approaches can be categorised broadly in three types. Detecting NER with predefined dictionaries and rules(R. Florian & Zhang, 2003), with some statistical approaches(Ratinov & Roth, 2009) and with deep learning approaches(X. Dong & Yang, 2016).

Stanford CoreNLP NER is a widely used baseline for many applications (Christopher Manning & McClosky, 2014). Authors have used approaches of Gibbs sampling and conditional random field (CRF) for non-local information gathering and then Viterbi algorithm to infer the most likely state in the CRF sequence output(Jenny Rose Finkel & Manning, 2005).

Deep learning approaches in NLP use document, word or token representations instead of one-hot encoded vec-

tors. With the rise of transfer learning, pretrained Word2Vec(Mikolov et al., 2013), GloVe(Jeffrey Pennington & Manning., 2014), fasttext(Piotr Bojanowski & Mikolov., 2017) which provides word embeddings were being used with recurrent neural networks (RNN) to detect NERs. Using LSTM layers followed by CRF layers with pretrained word-embeddings as input has been explored here(Huang et al., 2015). Also, CNNs with character embeddings as inputs followed by bi-directional LSTM and CRF layers, were explored here(Ma & Hovy, 2016).

With the introduction of attentions and transformers(Vaswani et al., 2017) many deep architectures emerged in last few years. Approach of using these pretrained models like Elmo(Peters et al., 2018), Flair(Akbik et al., 2019) and BERT(Devlin et al., 2018) for word representations followed by variety of LSMT and CRF combinations were tested by authors in (Straková et al., 2019) and these approaches show state-of-the-art performance.

There are very few approaches where caseless NER task is explored. In this recent paper(Mayhew et al., 2019b) authors have explored effects of "Cased" entities and how variety of networks perform and they show that the most effective strategy is a concatenation of cased and lowercased training data, producing a single model with high performance on both cased and uncased text.

In another paper(Mayhew et al., 2019a), authors have proposed True-Case pre-training before using BiLSTM+CRF approach to detect NERs effectively. Though it shows good results over previous approaches, it is not useful in Indian Languages context as there is no concept of cases.

In our approach, we are focusing more on data preparation for our definition of topics using some of the state-of-art architectures based on BERT, LSTM/GRU and CRF layers as they have been explored in previous approaches mentioned above. Detecting caseless topics with higher recall and reasonable precision has been given a priority over f1 score. And comparisons have been made with available and ready-to-use open-source libraries from the productionization perspective.

¹<https://www.dailyhunt.com>

²<https://github.com/flairNLP/flair>

³<https://spacy.io/api/entityrecognizer>

⁴<https://stanfordnlp.github.io/CoreNLP/caseless.html>

2. Data Preparation

We need good amount of data to try deep learning state-of-the-art algorithms. There are lot of open datasets⁵ available for names, locations, organisations, but not for topics as defined in Abstract above. Also defining and inferring topics is an individual preference and there are no fix set of rules for its definition. But according to our definition, we can use wikipedia titles as our target topics. English wikipedia dataset⁶ has more than 18 million titles if we consider all versions of them till now. We had to clean up the titles to remove junk titles as wikipedia title almost contains all the words we use daily. To remove such titles, we deployed simple rules as follows -

- Remove titles with common words : "are", "the", "which"
- Remove titles with numeric values : 29, 101
- Remove titles with technical components, driver names, transistor names : X00, lga-775
- Remove 1-gram titles except locations (almost 80% of these also appear in remaining n-gram titles)

After doing some more cleaning we were left with 10 million titles. We have a dump of 15 million English news articles published in past 4 years. Further, we reduced number of articles by removing duplicate and near similar articles. We used our pre-trained doc2vec models and cosine similarity to detect almost similar news articles. Then selected minimum articles required to cover all possible 2-grams to 5-grams. This step is done to save some training time without loosing accuracy. Do note that, in future we are planning to use whole dataset and hope to see gains in F1 and Recall further. But as per manual inspection, our dataset contains enough variations of sentences with rich vocabulary which contains names of celebrities, politicians, local authorities, national/local organisations and almost all locations, India and International, mentioned in the news text, in last 4 years.

We then created a parallel corpus format as shown in Table 1. Using pre-trained Bert-Tokenizer⁷ from hugging-face, converted words in sentences to tokenes. Caseless-BERT pre-trained tokenizer is used. Notice that some of the topic words are broken into tokens and *NER* tag has been repeated accordingly. For example, in Table 1 second row, word "harassment" is broken into "har ##ass ##ment". Similarly, one "NER" tag is repeated three times to keep the length

⁵<https://github.com/juand-r/entity-recognition-datasets>

⁶<https://dumps.wikimedia.org/enwiki/>

⁷<https://pypi.org/project/pytorch-pretrained-bert/>

Table 1. Parallel Corpus Preparation with BERT Tokenizer

Text	the me too movement with a large variety of local and international related names , is a movement against sexual harassment and sexual assault
NER Tags	0 NER NER NER 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 NER NER 0 NER NER
Tokenized Text	the me too movement with a large variety of local and international related names, is a movement against sexual har ##ass ##ment and sexual assault
Tokenized NER Tags	0 NER NER NER 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 NER NER NER NER 0 NER NER

of sequence-pair same. Finally, for around 3 million news articles, parallel corpus is created, which is of around 150 million sentences, with around 3 billion words (all lower cased) and with around 5 billion tokens approximately.

3. Experiments

3.1. Model Architecture

We tried multiple variations of LSTM and GRU layes, with/without CRF layer. There is a marginal gain in using GRU layers over LSTM. Also, we saw gain in using just one layers of GRU instead of more. Finally, we settled on the architecture, shown in Figure 1 for the final training, based on validation set scores with sample training set.

Text had to be tokenized using pytorch-pretrained-bert as explained above before passing to the network. Architecture is built using tensorflow/keras. Coding inspiration taken from BERT-keras⁸ and for CRF layer keras-contrib⁹. If one is more comfortable in pytorch there are many examples available on github, but pytorch-bert-crf-ner¹⁰ is better for an easy start.

We used BERT-Multilingual model so that we can train and fine-tune the same model for other Indian languages. You can take BERT-base or BERT-large for better performance with only English dataset. Or you can use DistilBERT for English and DistilMBERT for 104 languages¹¹ for faster pre-training and inferences. Also, we did not choose AutoML approach for hyper-parameter tuning which could have resulted in much more accurate results but at the same

⁸<https://github.com/Separius/BERT-keras>

⁹<https://github.com/keras-team/keras-contrib/>

¹⁰<https://github.com/eagle705/pytorch-bert-crf-ner>

¹¹<https://github.com/huggingface/transformers/tree/master/examples/distillation>

time could have taken very long time as well. So instead, chose and tweaked the parameters based on initial results.

We trained two models, one with sequence length 512 to capture document level important n-grams and second with sequence length 64 to capture sentence/paragraph level important n-grams. Through experiments it was evident that, sequence length plays a vital role in deciding context and locally/globally important n-grams. Final output is a concatenation of both the model outputs.

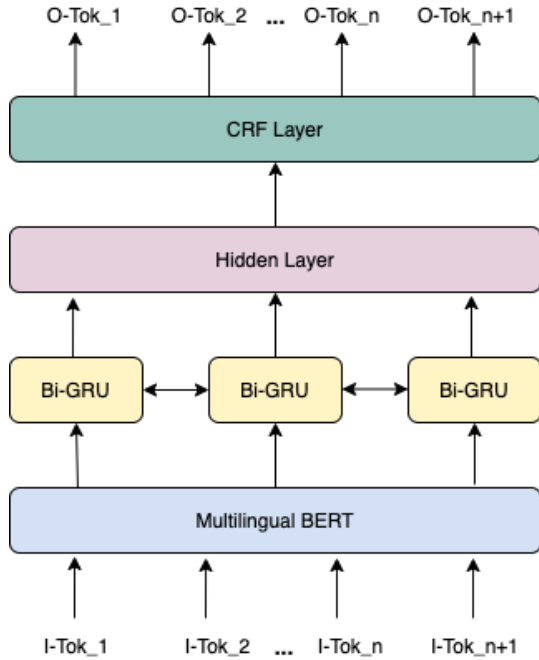


Figure 1. BERT + Bi-GRU + CRF, Final Architecture Chosen For Topic Detection Task.

3.2. Training

Trained the topic model on single 32gb NVidia-V100 and it took around 50 hours to train the model with sequence length 512. We had to take 256gb ram machine to accommodate all data in memory for faster read/write. Also, trained model with 64 sequence length in around 17 hours.

It is very important to note that sequence length decides how many bert-tokens you can pass for inference and also decides training time and accuracy. Ideally more is better because inference would be faster as well. For 64 sequence length, we are moving 64-token window over whole token-text and recognising topics in each window. So, one should choose sequence length according to their use case. Also, we have explained before our motivation of choosing 2 separate sequence lengths models.

We stopped the training for both the models when it crossed 70% precision, 90% recall on training and testing sets, as we

Table 2. Comparison with Traditional NERs as reference

Models	Precision	Recall	F1
BERT+BiGRU+CRF	60.09	80.08	68.66
Stanford	90.54	37.17	52.70
Spacy	88.71	55.05	67.94
Flair	85.62	10.28	18.36

Table 3. Comparison with Wikipedia titles as reference

Models	Precision	Recall	F1
BERT+BiGRU+CRF	51.97	69.76	59.56
Stanford	52.83	19.88	28.89
Spacy	36.31	26.40	30.57
Flair	65.36	7.33	13.17

were just looking to get maximum recall and not bothered about precision in our case. Both the models reach this point at around 16 epochs.

3.3. Results

Comparison with existing open-source NER libraries is not exactly fair as they are NOT trained for detecting topics and important n-grams, also NOT trained for case-less text. But they are useful in testing and benchmarking if our model is detecting traditional NERs or not, which it should capture, as Wikipedia titles contains almost all Names, Places and Organisation names. You can check the sample output here¹²

Comparisons have been made among Flair-NER, Stanford-caseless-NER (used english.conll.4class.caseless as it performed better than 3class and 7class), Spacy-NER and our models. Of which only Stanford-NER provides case-less models. In Table 2, scores are calculated by taking traditional NER list as reference. In Table 4, same is done with Wikipedia Titles reference set.

As you can see in Table 2 & 3, recall is great for our model but precision is not good as Model is also trying to detect new potential topics which are not there even in reference Wikipedia-Titles and NER sets. In capturing Wikipedia topics our model clearly surpasses other models in all scores.

Spacy results are good despite not being trained for case-less data. In terms of F1 and overall stability Spacy did better than Stanford NER, on our News Validation set. Similarly, Stanford did well in Precision but could not catch up with Spacy and our model in terms of Recall. Flair overall performed poorly, but as said before these open-source models are not trained for our particular use-case.

¹²https://github.com/swapniljadhav1921/bert_crf_topic_detection/

3.4. Discussions

Lets check some examples for detailed analysis of the models and their results. Following is the economy related news.

Example 1 : *around \$1–1.5 trillion or around two percent of global gdp, are lost to corruption every year, president of the natural resource governance institute nrgi has said. speaking at a panel on integrity in public governance during the world bank group and international monetary fund annual meeting on sunday, daniel kaufmann, president of nrgi, presented the statistic, result of a study by the nrgi, an independent, non-profit organisation based in new york. however, according to kaufmann, the figure is only the direct costs of corruption as it does not factor in the opportunities lost on innovation and productivity, xinhua news agency reported. a country that addresses corruption and significantly improves rule of law can expect a huge increase in per capita income in the long run, the study showed. it will also see similar gains in reducing infant mortality and improving education, said kaufmann.*

Detected NERs can be seen per model in Table 4. Our model do not capture numbers as we have removed all numbers from my wiki-titles as topics. Reason behind the same is that we can easily write regex to detect currency, prices, time, date and deep learning is not required for the same. Following are few important n-grams only our models was able to capture -

capita income
infant mortality
international monetary fund annual meeting
natural resource governance institute
public governance

At the same time, we can see that Spacy did much better than Stanford-caseless NER and Flair could not capture any of the NERs. Another example of a news in political domain and detected NERs can be seen per model in Table 5.

Example 2 : *wearing the aam aadmi party's trademark cap and with copies of the party's five-year report card in hand, sunita kejrwal appears completely at ease. it's a cold winter afternoon in delhi, as the former indian revenue service (irs) officer hits the campaign trail to support her husband and batchmate, chief minister arvind kejrwal. emerging from the background for the first time, she is lending her shoulder to the aap bandwagon in the new delhi assembly constituency from where the cm, then a political novice, had emerged as the giant killer by defeating congress incumbent sheila dikshit in 2013.*

Correct n-grams captured only by our model are -

aam aadmi party
aap bandwagon
delhi assembly constituency
giant killer
indian revenue service
political novice

In this example, Stanford model did better and captured names properly, for example "*sheila dikshit*" which Spacy could not detect but Spacy captureed almost all numeric values along with numbers expressed in words.

It is important to note that, our model captures NERs with some additional words around them. For example, "president of nrgi" is detected by the model but not "nrgi". But model output does convey more information than the later. To capture the same for all models (and to make comparison fair), partial match has been enabled and if correct NER is part of predicted NER then later one is marked as matched. This could be the reason for good score for Spacy. Note that, partial match is disabled for Wikipedia Titles match task as shown in Table 3. Here, our model outperformed all the models.

4. Conclusion and Future Work

Through this exercise, we were able to test out the best suitable model architecture and data preparation steps so that similar models could be trained for Indian languages. Building cased or caseless NERs for English was not the final goal and this has already been benchmarked and explored before in previous approaches explained in "Related Work" section. We didn't use traditional datasets for model performance comparisons & benchmarks. As mentioned before, all the comparisons are being done with open-source models and libraries from the productionization point of view. We used a english-news validation dataset which is important and relevant to our specific task and all validation datasets and raw output results can be found at our github link ¹³.

Wikipedia titles for Indian languages are very very less and resulting tagged data is even less to run deep architectures. We are trying out translations/transliterations of the English-Wiki-Titles to improve Indic-languages entity/topics data.

This approach is also useful in building news-summarizing models as it detects almost all important n-grams present in the news. Output of this model can be introduced in a summarization network to add more bias towards important words and bias for their inclusion.

References

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., and Vollgraf, R. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 54–59, Minneapolis, Minnesota, June 2019. Asso-

¹³https://github.com/swapniljadhav1921/bert_crf_topic_detection

- ciation for Computational Linguistics. doi: 10.18653/v1/N19-4010. URL <https://www.aclweb.org/anthology/N19-4010>.
- Christopher Manning, Mihai Surdeanu, J. B. J. F. S. B. and McClosky, D. The stanford corenlp natural language processing toolkit. *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 2014.*, 2014.
- Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv e-prints*, art. arXiv:1412.3555, Dec 2014.
- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
- Huang, Z., Xu, W., and Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv e-prints*, art. arXiv:1508.01991, Aug 2015.
- Jeffrey Pennington, R. S. and Manning., C. Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
- Jenny Rose Finkel, T. G. and Manning, C. Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics - ACL '05*, 2005.
- Ma, X. and Hovy, E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *arXiv e-prints*, art. arXiv:1603.01354, Mar 2016.
- Mayhew, S., Gupta, N., and Roth, D. Robust Named Entity Recognition with Truecasing Pretraining. *arXiv e-prints*, art. arXiv:1912.07095, Dec 2019a.
- Mayhew, S., Tsygankova, T., and Roth, D. ner and pos when nothing is capitalized. *arXiv e-prints*, art. arXiv:1903.11222, Mar 2019b.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv e-prints*, art. arXiv:1301.3781, Jan 2013.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. Deep contextualized word representations. *arXiv e-prints*, art. arXiv:1802.05365, Feb 2018.
- Piotr Bojanowski, Edouard Grave, A. J. and Mikolov., T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.
- R. Florian, A. Ittycheriah, H. J. and Zhang, T. Named entity recognition through classifier combination. *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4, 2003: Association for Computational Linguistics*, 2003.
- Ratinov, L. and Roth, D. Design challenges and misconceptions in named entity recognition. *Proceedings of the thirteenth conference on computational natural language learning, 2009: Association for Computational Linguistics*, 2009.
- Straková, J., Straka, M., and Hajič, J. Neural Architectures for Nested NER through Linearization. *arXiv e-prints*, art. arXiv:1908.06926, Aug 2019.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention Is All You Need. *arXiv e-prints*, art. arXiv:1706.03762, Jun 2017.
- X. Dong, L. Qian, Y. G. L. H. Q. Y. and Yang, J. A multiclass classification method based on deep learning for named entity recognition in electronic medical records. *New York Scientific Data Summit (NYSDS), 2016: IEEE*, 2016.

Flair	Spacy	Stanford	BERT+BiGRU+CRF
	\$1–1.5 trillion	daniel	around two percent
	annual	international monetary fund	bank
	around two percent	new york.	capita income
	daniel kaufmann	ngi	daniel kaufmann
	every year	xinhua	every year
	kaufmann		infant mortality
	new york		international monetary fund annual meeting
	ngi		natural resource governance
	sunday		natural resource governance institute
	the natural resource governance institute nrgi		new york
	the world bank group		public governance
	xinhua news agency		rule
			the natural resource governance institute nrgi
			the world bank group
			xinhua news agency

Table 4. Recognised Named Entities Per Model - Example 1

Flair	Spacy	Stanford	BERT+BiGRU+CRF
indian	2013	aam aadmi	aam aadmi party
sheila dikshit	aap	arvind kejriwal.	aap bandwagon
	arvind kejriwal	congress	arvind kejriwal
	congress	ease.	delhi
	delhi	indian	delhi assembly constituency
	first	new delhi	giant killer
	five-year	sheila dikshit	indian revenue service
	indian	sunita kejriwal	political novice
	irs		sheila dikshit
	sunita kejriwal		sunita kejriwal
	the aam aadmi party's		the aam aadmi party
	winter afternoon		winter afternoon

Table 5. Recognised Named Entities Per Model - Example 2