Technology Extraction - NER

May - Aug 2021

—

SUMMER INTERNSHIP REPORT- 2021

Mohit Bagaria
17EX20017
Geology & Geophysics
Data Science Intern
Slintel

ACKNOWLEDGEMENT

I'd like to thank my department Geology & Geophysics and my faculty Advisor Dr. Arun Singh Sir, for guiding and molding me throughout the course of time.

I would like to express my gratitude to Slintel for giving me the opportunity to work on this project. It has been an exponential learning experience.

I would like to thank my Mentor and guide, Mr. Rishi Agarwal , for his constant guidance throughout the course of the project. Frequent virtual meetings and proper guidance helped me to head in the right direction and eventually we obtained quite satisfactory results.

Finally, I apologise for missing out on thanking all other unnamed people who helped me in various ways and kept me motivated throughout the course of the project.

INDEX

INTRODUCTION - SLINTEL

A Sales Intelligence Platform, which helps to Identify Potential buyers in their Market & uncover top 5% prospects in their segment, using recent data

Use Slintel to identify the best selling opportunities that you can reach out to today

Get verified emails and direct dials for active, high intent buyers in your target markets

Understand buyer behaviour and pain points using buyer journeys and keyword insights

Streamline your pitch with technology adoption data

All in one place- company, contact, technology and intent data

Problem Statement

1. Technologies Extraction From Linkedin Summaries
2. Data Available-

    Technologies Names

    Scraped Linkedin Profiles

Project objective

To extract the Organisation names and the Technological keywords mentioned in their LinkedIn About or Summary section

This Problem Statement is Termed as NAMED ENTITY RECOGNITION (NER)

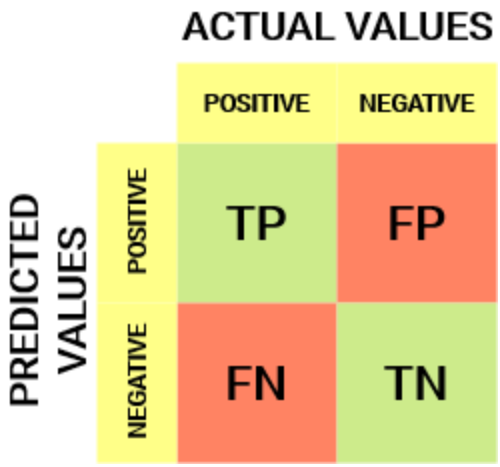
Pre-Requisites:

I. Tagging Data:

Each word has to be tagged using String Match and then with the help of Human Intervention he/she has to check contextually whether the word is in real referring to an Organization or Technology

II. Confusion Matrix - used for comparison among different approaches

A Confusion matrix is an N x N matrix used for evaluating the performance of a classification model, where N is the number of target classes. The matrix compares the actual target values with those predicted by the machine learning model. This gives us a holistic view of how well our classification model is performing and what kinds of errors it is making.



True Positive (TP) :

The predicted value matches the actual value

The actual value was positive and the model predicted a positive value

True Negative (TN) :

The predicted value matches the actual value

The actual value was negative and the model predicted a negative value

False Positive (FP): – Type 1 error

The predicted value was falsely predicted. The actual value was negative but the model predicted a positive value

False Negative (FN): – Type 2 error

The predicted value was falsely predicted. The actual value was positive but the model predicted a negative value

Precision:

Precision tells us how many of the correctly predicted cases actually turned out to be positive

(true positives/predicted positives)=TP/(TP+FP)

Recall:

Recall tells us how many of the actual positive cases we were able to predict correctly with our model.

(true positives/all actual positives)=TP/(TP+FN)

F Score :

F1-score is a harmonic mean of Precision and Recall, and so it gives a combined idea about these two metrics. It is maximum when Precision is equal to Recall. The interpretability of the F1-score is poor. This means that we don't know what our classifier is maximizing – precision or recall? So, we use it in combination with other evaluation metrics which gives us a complete picture of the result.

Harmonic Mean of Precision and Recall => 2*(precision*recall)/(precision+recall)

Range of Precision, Recall and F1 score is between (0,1)

MODEL CYCLE-

Step 0: Data Collection-

Slintel's Data Engineers had scraped three kinds of data

1. LinkedIn Summaries (About section) of People
2. Technology Dictionary(, ~ 24000 rows; Technological words for which we are sure are Technical or Organizational words, Ex= GitHub, MongoDB since these words have no other meaning other than a Technical one)
3. English Dictionary, ~6000 (Contextual words which might not be a technology , ex-'Python' can referred as a snake and also as a Programming Language, WorkSpace)

Step 1: Cleaning & Tagging-

I removed the unnecessary symbols like & - / , : ; which might not be relevant. Apart from this common words like is the can be removed since it is basically a Noise in our data .Converting all text to lowercase

Marking all the words using String Match which will be used to convert text into BIO Format

Then remove all the words which contextually meant something else,i.e., not referring to any Organization or Technology

Step 2: Conversion to BIO format

All sentences are broken into words separated by space and tagged as B-org or I-org & O (The B- prefix before a tag indicates that the tag is the beginning of a chunk, & an I- prefix before a tag indicates that the tag is inside a chunk. An O tag indicates that a token belongs to no entity / chunk.)

BIO Format example

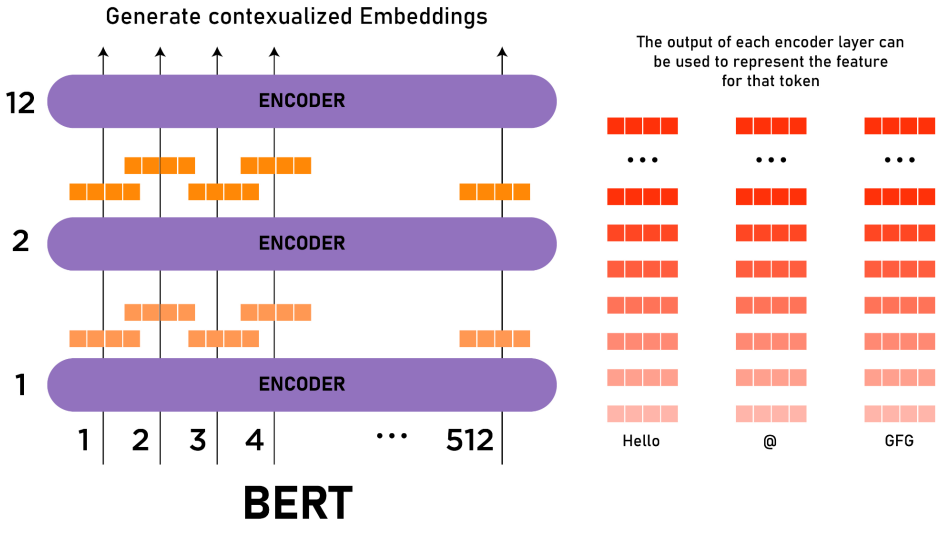
Text	Label
In	O
Beirut	B-geo
,	O
a	O
string	O
of	O
officials	O
voiced	O
their	O
anger	O
,	O
while	O
at	O
the	O
United	B-org
Nations	I-org
summit	O
in	O
New	B-geo
York	I-geo
,	O
Prime	B-per
Minister	O
Fouad	B-per
Siniora	I-per
said	O
the	O
Lebanese	B-gpe
people	O
are	O
resolute	O

Step 3: Fine Tuning BERT

BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers) is a recent paper published by researchers at Google AI Language. It has caused a stir in the Machine Learning community by presenting state-of-the-art results in a wide variety of NLP tasks, including Question Answering (SQuAD v1.1), Natural Language Inference (MNLI), etc.

BERT makes use of Transformer, an attention mechanism that learns contextual relations between words (or sub-words) in a text.

Loaded Google's pre-trained model with the help of Huggingface library and then Fine-Tuned Bidirectional Encoder Representations from Transformers (bert-ner-uncased) model on our data to generate future predictions from text.



RESULTS

FINAL SCORES:

PRECISION -0.85

RECALL - 0.83

F1_SCORE- 0.84

K-fold Cross Validation Stats: (k=5)

It is used to check if our Model is overfitting for any particular Data

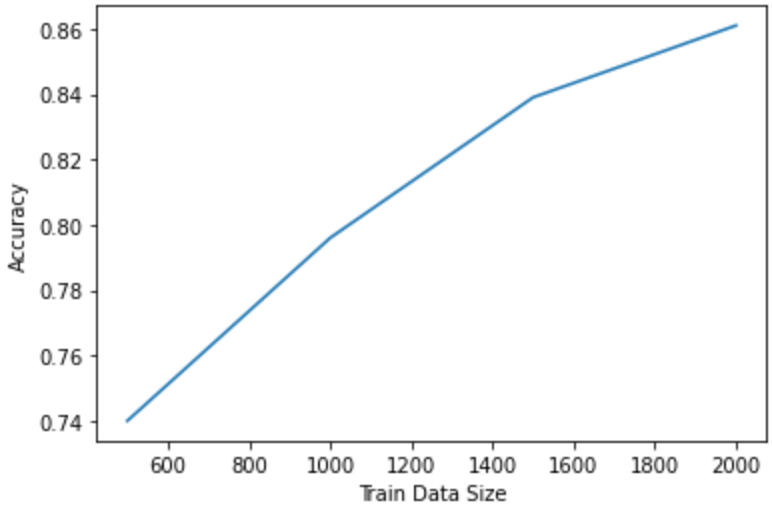
Resetting Model weights after each Training:

Training on 2000, testing on 500 rest UNSEEN data

TRAIN	TEST	Precision	Recall	F1-Score
1,2,3,4	5	0.844	0.847	0.846
1,2,3,5	4	0.838	0.840	0.839
1,2,4,5	3	0.853	0.831	0.842
1,3,4,5	2	0.857	0.837	0.847
2,3,4,5	1	0.835	0.845	0.840
	AVERAGE	0.8454	0.840	0.842

Variation with Train Data Size vs Accuracy:

TRAIN	TEST	Precision	Recall	F1-Score
1	5	0.75	0.72	0.74
1,2	5	0.803	0.789	0.796
1,2,3	5	0.847	0.83	0.839
1,2,3,4	5	0.860	0.863	0.861



Here, we can clearly conclude that Accuracy will increase with Increase in Train Data Size

Further Improvement:

I used an ENSEMBLE approach where we will use String Match on Technical Words(Non-Contextual) and BERT Predictions only on English words(Contextual words, which may or may not be a Technological word). This is our exact need and will further improve the Usability and Precision of the Model

ENSEMBLE MODEL = STRING MATCH + BERT ON ENGLISH(Contextual Dictionary words)

Approach	Precision	Recall	F1-Score
String Match	0.57	0.49	0.53
BERT	0.847	0.83	0.84
String Match on TECH | BERT on ENG	0.89	0.92	0.904

Thank You!

-By Mohit Bagaria

17EX20017

GEOLOGY & GEOPHYSICS

Indian Institute of Technology Kharagpur
Kharagpur 721302, India
CAREER DEVELOPMENT CENTRE
CERTIFICATE OF STUDENT'S PRACTICAL TRAINING

1. Name : Mohit Bagaria
2. Roll No. : 17EX20017
3. Year of Study : 4th
4. Branch &Department : Exploration Geophysics, Geology & Geophysics
5. Name and Address of Organization : Slintel, Bangalore, India.

6. Place of Training : Remote
7. Date of Commencement of Training : 10th May, 2021
8. Date of Completion of Training : 9th August, 2021
9. Number of Working Days Attended : 64
10. Days of Leave Availed, if any : 0
11. Overall Performance of the Student during Training:

Excellent [✓] Good [ ] Satisfactory [ ] Unsatisfactory [ ]

12. The work carried out here contain confidential data YES [✓] NO [ ]

If YES, please fill the additional **confidentiality disclaimer**

Remarks on the conduct of the Student, Punctuality and Interest etc.:

.....Mohit was very sincere with regards to his work and always ensured to complete all the allocated tasks on time.....

Date: Signature of the Authorized Officer.....

Name & Designation of the Officer (with Seal) Rahul Bhattacharya, Co-Founder & CTO

Note: Student should obtain 3 copies of this, one for the Organization, one for CDC and the other to be included in the final report to be submitted to the department.

CONFIDENTIALITY DISCLAIMER

Data Confidentiality Statement:

The work carried out at Slintel India Private Limited contains confidential data that cannot be used for the report purposes of the student. However, the process and methodology details can be included in the report.

Mohit Bagaria

Company Official Signature **Student Signature**