A Review of Named Entity Recognition

Ms. Komal Domadiya[1], Mr. Hetul Patel[2]

[1,2]Babu Madhav Institute of Information Technology, Uka Tarsadia University, Bardoli, Surat, Gujarat, India

---***---

Abstract - *Named entity recognition is one of the major tasks under Natural Language processing (NLP), which is widely used in the fields of computer science and software development social sites. such as a Q&A on a stack overflow, Quora or CSDN. They have reach information of software functions. Social sites have some difficulties to the SFF (software function feature) specific NER with the BI-LSTM(Bidirectional-long short-term memory) model. Existing approaches cannot support to the direct answer & knowledge graph. Existing NER methods are designed for recognize person, location informal and social texts, which are not applicable to NER software engineering. Our NER system is called S-NER(software named entity recognition).*

Key Words: Named entity recognition, stack overflow, Quora, CSDN, BI-LSTN, CRF, SFF, S-NER.

1. INTRODUCTION

The term NER represents one of the most intensively studied information extraction tasks. This task is defined as the task of finding structured information for unstructured or semi-structured text.[4] NER is first used at the sixth message understanding conference (MUC-6) as the task of identifying names of person, locations and organizations in text [11]. But such as many problems in the machine translation and Question answering etc [12]. In the Q&A system receives the users questions then it first select a set of relevant documents, and then filters out irrelevant pieces of text of these documents gradually until the answer is found [13]. But In today's era Social development sites have become more popular because of the rapid development of internet and huge amount of information. Social sites such as a stack overflow and Quora play a significant role in knowledge sharing and acquisition for software developer.

A fundamental task for reusing content in this website is searching for discussion of specific software entity (e.g. library, tool, an API) to find bugs and alternatives. Existing approaches are using textual documents and use vector space model [5].

A neural network model is considered to be a less feature. So however neural networks have some limitations such as a relying on simple feed forward for neurons learning and depending on word embedding text features [10]. So, we use the two models.

There are Two types of Model: -

1. **CRF model (Conditional random field)**
2. **BI-LSTM (Bidirectional-long short-term memory)**

The question & answers are all represented by natural languages in software development social networking sites. Over 92% of stack overflow questions about expert topics are answered in time of 11 minutes [8]. There is no appropriate recognition approach for SFF [6].

When writing a software function, question or blog, the user describes the software function feature as their own way. Moreover, software function feature data is usually Chinese or English words. It makes the SFF specific NER more difficult [6].

NER has been extensively studied on formal text which are recognize the real-world objects in text such as person, location or organization name [1]. Existing approaches are limited to dictionary look-up, furthermore the entity category is limited to only API. In this work we design and evaluate a machine learning based for general NER in software engineering social content. In this NER we would like to recognize not only API but other categories of software specific entities such as programming languages, platforms, tools, libraries, frameworks, software standards.

Machine learning based NER method is called S-NER, including software specific entity category, a software-specific tokenizer, conditional random field (CRF) based learning. And a reach and effective set of features for model training [5]. Our models rely on two sources of information about words: character-based word representation learned from the supervised corpus and unsupervised word representation learned from unannotated corpora.

Table.1-Software-specific entity Categories [6].

Entity category	Anno.Tag	Examples
Programming languages	PL	• Object-oriented-java,C#
Platform	Plat	• Cpu

		instruction sets
API	API	• Java-script onclick event
Tool library framework	Fram	• Software tools
Software standards	Stan	• Ajax,JDBC

3. CRF model (Conditional random field)

CRF model is a class of statistical modelling method often applied in pattern recognition and machine learning and used for structured prediction. What kind of graph is used in the application for example in natural language processing, linear chain CRF is popular which implement sequential dependencies in the predictions. In image processing the graph typically connects locations to nearby or similar location to enforce that they receive similar predictions.

Conditional Random fields is an undirected graph model in which each word in the input sentence is represented using a corresponding tag in the output sentence. The CRF model depending on the neighbouring words/tags to predict the tag of the current word, the prediction of the tag can be done using a sentence level presentation as a whole instead of the individual word representation.

The CRF model was trained to label three categories: person, location, and organization and for other categories are Object [10].

Table.2: - NER Used to train our Models [6].

Main class	Sub class
Person(PER)	-Politician -Scientist -Business -Artists -Police -Group -Engineer
Organization(ORG)	-Government -education -Media -commercial -sports -religious
Location (Loc)	-Water body -Celestial -Land region nature

3.1 Design challenges in NER in software engineering social content.

As we know existing NER methods can recognize only real-world objects. So, recognize software specific entities we must develop software specific NER methods but excepts some API extraction and linking work in software engineering text [5].

We randomly sample a diverse 150 stack overflow posts covering 6 popular programming languages such as java script, java, C#, python, PHP, Html and also 1 popular platform android and jQuery. Then we manually identify software specific entities in these sample posts. Through this formative study.

3.2 Challenges IN NER

In the stack overflow discussion, more spelling mistakes are made compare to formal texts. For example, capitalization is used to extensively in question titles and discussions for emphasis It happens that “JavaScript” is misspelled as “JavaScript” (missing the character C) [5].

Different software entities often have the same name. For example, the term “Memcached” can be a PHP class. And can also be a memory management tool.” Mac” can be a platform name or class name of android. This causes ambiguity in determining appropriate entity category.

The informal nature of Stack overflow posts introduces many name variations for the same software specific entity. For example, in addition to official programming language JavaScript also refer to the languages as java script or js or JS.

Challenges 1 and 2 indicates that dictionary look up or rule-based methods would not produce reliable NER results on software engineering social content. Rule based approaches are typically done using handcrafted rules and gazetteers lists. The drawbacks of rule-based approaches are that they require linguistics expertise and are domain specific [3].

Challenge 3 is indicating that stack overflow discussions contain many more out-of-Vocabulary (OOV) words (i.e. entities that have not been seen in the training data.)

Due to OOV words it is difficult to identify using local contextual cues alone because the entity has not been seen before [5].

3.3 The software specific NER System

S-NER is based on conditional random fields (CRF) for supervised model training. In this section we discuss data preparation steps, customized tokenization, unsupervised word clustering.

Data preparation

1. Labelled Data Preparation.
2. Unlabelled Data Preparation.
3. External Knowledge Resource

- **Labelled Data Preparation: -**

We randomly select posts under a diverse set of stack overflow tags, representing popular object oriented and procedural languages (java, C#) web & scripting languages, mark-up languages, platform and library (jQuery). In fact, these 8 tags are most frequently used tags among all the stack overflow tags. We select 1520 stack overflow posts from 300 stack overflow discussion threads. We refer to this dataset as Labelled data.

- **Unlabelled Data Preparation**: -

We randomly select a huge sized data consisting of more than 7 million stack overflow posts from 1.8 million stack overflow discussion threads tagged with the 8 most frequently used tags such as java, java script, C#, python, html, android, PHP and jQuery. We refer to this dataset as unlabelled data.

- **External Knowledge Resource: -**

We can use gazetteers as features for training a CRF model. There are many gazetteers publicly available for common person names, location, organizations and products. That can help to recognize software specific entities in software engineering texts [5].

It has been known that gazetteers or entity dictionaries are important for the improving the performance of named entity recognition [9].

Figure1: - S-NER System working Flow. [5]

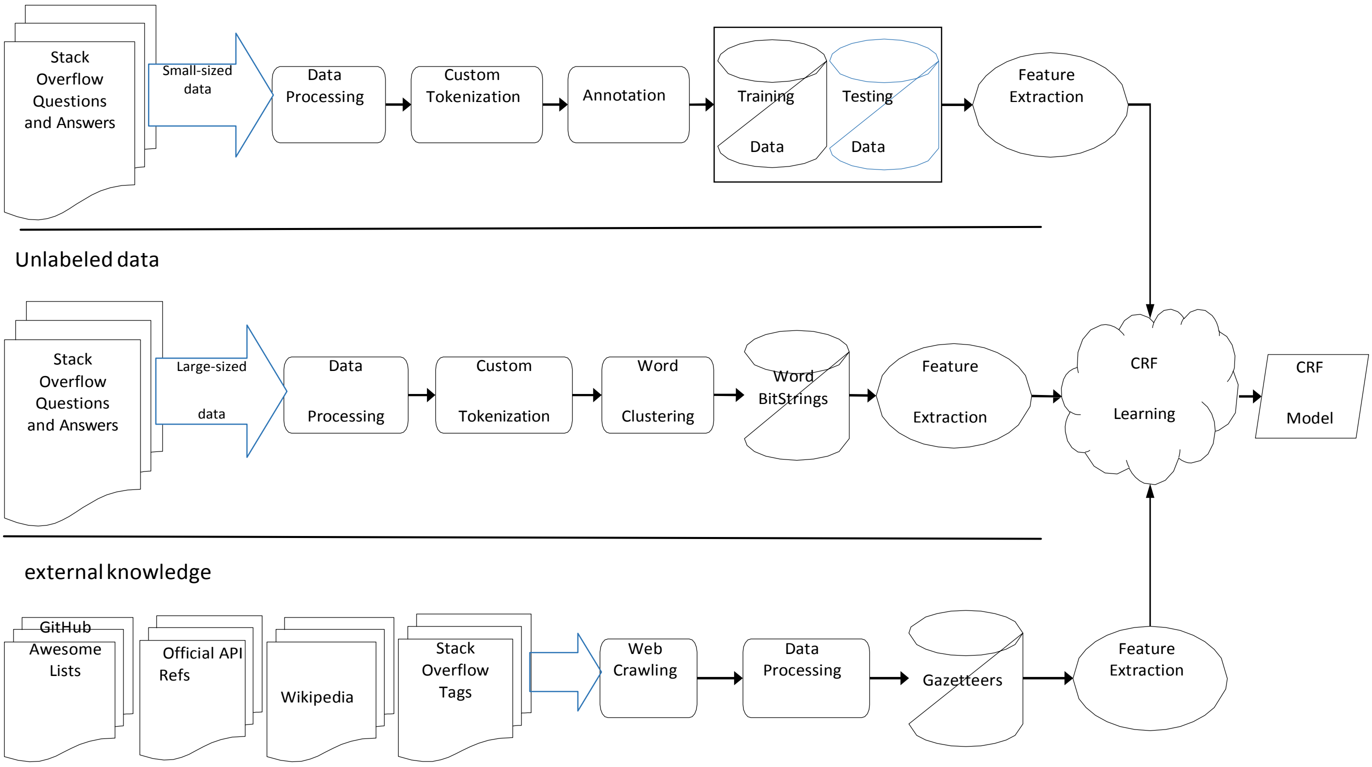


Table3: - Example of S-NER tokenization [5].

Input Sentence	What's the equivalent of java's Thread.sleep() in js?
Stanford Tokenizer	What 's the equivalent of java's Thread. sleep ( ) in js ?
S-NER Tokenization	What's the equivalent of java's Thread.sleep() in js ?

❖ Customized Tokenization:

We develop a domain-specific tokenizer for handling text with software specific entities. The tokenizer does not split The name of software entity. E.g. the name of an API it does not split valid programming operators such as '==' or '!=' it consider separate parentheses. However, parentheses as well as dot, # and $, that appear in an API are consider as part of API itself. In table 3 we show an example of the S-NER tokenization results.

Line 1 is the input sentence, and line 2 is the tokenization done by stand ford Tokenizer.

❖ Unsupervised words clustering

The problem of OOV (out of vocabulary) and lexical word variations, we rely on unsupervised word clustering to group together words that are distributional similar, we apply Brown clustering for unlabelled stack overflow posts. We use liang's implementation of brown clustering we configure the number of clusters to 1000 and we only cluster word that appear no less than 10 times. It takes 15 hours to complete word clustering on the unlabelled dataset on a 4-Core intel i5-4570 processor.

Table4: -Example of word clustering results [5].

Bitstring	Top words (by frequency)
1011110	.NET Spring ASP.NET Django HTML5 asp.net bootstrap django Bootstrap Entity spring .Net .net wordpress Wordpress Angular AngularJS JPA
11111111110	Foreach setTimeout setInterval eval Files json encode_explode exec var dump print r await document. Write
111111111111111	hover touch mouseover blur keyup keypress key down mouseout fadeIn mouseenter mouseleave mousedown delegated show() fadeOut mouseup mousemove hide()

4. Bi-LSTM Model

To deal with challenges, we propose an approach for SFF specific named entity recognition with the BI-LSTM model and word embedding technique. We use Bidirectional long short-term memory (BI-LSTM) model. To induce character level feature, we use convolutional neural network which has been successfully applied to Spanish and Portuguese NER and German pos tagging [7].

The first work to recognize the SFF-specific named entity in software development social contents.

The rest of this paper is organized as follows section VI introduces the problem definition of SFF specific NER in software development social contents. Section VII describes the details of our approach [6].

4.1 Problem Definition

Based on our formative analysis, we define the SFF-specific NER problem in software development social networking sites contents. The sentence consists of many words, In which some serve as subject, other word serve as object and the rest serve as verbs, adverbials or complements. The words serving as subject or object are the entity of the sentence.

The entity of the software function feature specific consists of one or more words. So, when recognize SFF specific entities there are two situations.

The first one is SFF specific entity consists of one word. We just need to define if the word is an entity. There many ways to judge if the word is an entity. The simplest way is to query the dictionary for the existence of the word. If the word in the dictionary, we regard the word as an entity. There are many entities that consists of one word such as "GPU", "CPU", or "JAVA". In this situation, the boundary of the SFF specific entity is the word.

The second one is the SFF specific entity consists of multiple words. We need to judge the first word and the end of the word of the SFF-specific entity. Because the SFF specific entities are made up of many OOV words that are not in the training data or in the dictionary. For this situation we need to use the deep learning technology to judge the boundary of the SFF specific entity. Using a dictionary may define additional feature in the CRF model that represents the dependencies between the word's NER label.

In SFF-specific some entities can be function names which can be defined by developers themselves define or already exists in the tool library. In the software development social networking sites contents the number of function name as an entity is very high. In previous studies, researchers ignored the names of the functions. The function name in SFF-specific sentences serve as subject and we should recognize the function names as SFF-specific entities. Because the function names as an entity will help us to understand the semantics of the sentence. So, we add a category for these function name.

4.2 Approach

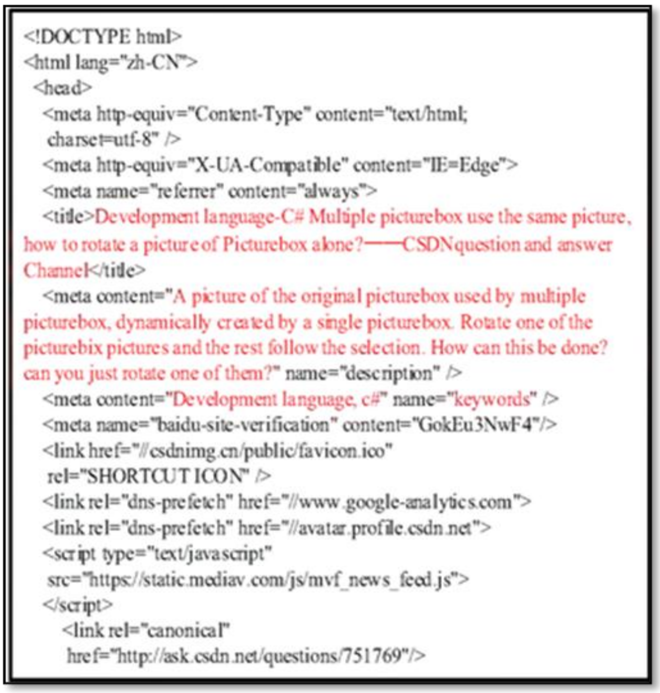
The system includes three steps: the first step is to collect and clean the data from CSDN by WebCrawler to get the text which doesn't contain the code and links. The second steps is to process the text into a short sentence which contains the SFF specific, segment the short sentence into words. Train the words to get the word embedding by word2vec. The third step is to training the BI-LSTM model to recognize the boundary of the SFF-specific entity and save the output results.

A. Data collection and cleaning

There are no studies in the NER for the function feature specific text in CSDN, so we firstly need to collect HTML pages from CSDN by a web crawler. CSDN is a Chinese stack overflow.

For these question and answer page source code, we have to analyse the structure of the page, find the position of tags, titles, question description and answers, extract this information and save this information into text format.

Figure2: - Data collection [6]



```
<!DOCTYPE html>
<html lang="zh-CN">
 <head>
  <meta http-equiv="Content-Type" content="text/html;
   charset=utf-8" />
  <meta http-equiv="X-UA-Compatible" content="IE=Edge">
  <meta name="referrer" content="always">
  <title>Development language-C# Multiple picturebox use the same picture,
how to rotate a picture of Picturebox alone?——CSDNquestion and answer
Channel</title>
  <meta content="A picture of the original picturebox used by multiple
picturebox, dynamically created by a single picturebox. Rotate one of the
picturebix pictures and the rest follow the selection. How can this be done?
can you just rotate one of them?" name="description" />
  <meta content="Development language, c#" name="keywords" />
  <meta name="baidu-site-verification" content="GokEu3NwF4"/>
  <link href="//csdnimg.cn/public/favicon.ico"
   rel="SHORTCUT ICON" />
  <link rel="dns-prefetch" href="//www.google-analytics.com">
  <link rel="dns-prefetch" href="//avatar.profile.csdn.net">
  <script type="text/javascript"
   src="https://static.mediav.com/js/mvf_news_feed.js">
  </script>
    <link rel="canonical"
     href="http://ask.csdn.net/questions/751769"/>
```

B. Word2vec for the word embedding

Word embedding refers to the mapping of chars or words in natural languages to low dimension dense real number vector, which is widely used to capture the words semantic and the sentence syntactic properties from natural language training data. In the word embedding the similarity of the word is encoded into the low dimension dense real number vector. So, if we want know the similarity between two or more words, we can calculate distance between these words vector in word vector space. For example, if we want to calculate similarity between four words Beijing, china, Seoul and Korea. The first is the vector of china minus the vector of Beijing. The second is the first result vector plus the vector of Seoul. The final result of vector is closest to the vector of Korea. It improves the performance of NLP tasks such as a named entity recognition, sentiment analysis or parsing [6].

❖ LITERATURE REVIEW

In the NER main problem is identifying the object whether it is a person name, organization or location, time or currencies [1].

Named entity recognition (NER) is used in the field of computer science and linguistics. And the drawbacks of rule-based approaches are they require linguistics expertise and are domain specific [3].

NER main task is finding the structured information from unstructured or semi-structured text [4].

NER is used in the social content Q&A sites such as a stack overflow or QUORA. Existing approaches are limited to only API and dictionary look up. But in this work, we evaluate a machine learning based method for NER and recognize SFF-specific NER such as: -

- Programming Languages
- Platforms
- Tools
- Libraries
- Frameworks
- Software standards [5].

When user is writing the software function, blog or any question as their own way. So, deal with this type of challenges we used the BI-LSTM model.

In the BI-LSTM model the system has two steps: -

1. Data collection and cleaning
2. Word2vec for word [6].

BI-LSTM model is used to induce character level feature. Which has been successfully applied to German POS Tagging [7].

Analyse question & answer dites for programmers stack overflow is over 92% questions about experts topics are answered in 11 minutes [8].

In the data preparation we use the external knowledge resource and it's important for improving the performance of named entity recognition gazetteers are very useful [9].

In existing approaches used the Neural network but the neural network has some limitations such as relying on simple feed forward for neurons learning and text features [10].

The term NER is first used in the MUC-6 (message understanding conference-6). That time it's used only for recognize person, location or organization name but later then it's used in the social Q&A sites and much more [11].

Previous NER is identifying the person name, location, organization etc. but many problems are occurred in such as Q&A and machine translation etc [12].

In the Q&A social content sites how, questions are accepted and reply to its answer [13].

❖ Acknowledgement

This work was supported by the respected guide Ms. Komal Domadiya from Uka Tarsadiya University (BMIIT).

❖ References

1. Adli Varlik tanimada yeni bir yaklasim A New approach for Named entity recognition, May 2017.

2. Ananya-A named entity recognition for Sinhala language In mortuva engineering research conference, April 2016.

3. A named entity recognisition approach for Albanian prof. marjana prifti skenduli, marenglen biba department of computer science university of new York in Tirana.marjanaprifti@unyt.edu.al.

4. Software specific NER in Software engineering Social content, Dehenge ye, jing li and nachiket kapre nanyang technological university Singapore. March 2016.

5. Feature specific NER in software development social content, authors: - ning li, liwei zheng, ying wang university of information science and technology Beijing china. August 2019.

6. Named entity recognition with bidirectional LSTM-CNN,author:- Jason pc.chiu university of british Columbia, eric Nichols Honda research institute japan.2016.

7. https://people.eecs.berkeley.edu/~bjoern/papers/mamykina-stackoverflow-chi2011.pdf

8. https://www.researchgate.net/publication/221013227_Exploiting_Wikipedia_as_External_Knowledge_for_Named_Entity_Recognition.

9. Using Bidirectional Long Short-Term Memory and Conditional Random Fields for Labeling Arabic Named Entities: A Comparative Study, author: - Sa'ad alzboun, saja khaled, mohammad AL-Smadi, Yaser jararweh. Jordan university of science & technology. October 2018.

10. A survey of deep learning named entity recognition https://www.researchgate.net/publication/329946134_A_Survey_on_Deep_Learning_for_Named_Entity_Recognition.

11. Bidirectional LSTM-CNN with extended features for named entity recognition. Author: - Necva bolucu, derman akgol, salih tuc. Hacettepe university, turkey April 2019.

12. Named entity recognition for question answering. researchgate.net/publication/228372089_Named_entity_recognition_for_question_answering.

13. https://arxiv.org/abs/1603.01360

14. https://en.wikipedia.org/wiki/Conditional_random_field