**CE807-7-SP: Text Analytics**

**Assignment 1**

**Literature Review**

**Name: Yaowei Lyu**

**Tutor: Dr habil Ansgar Scherp**

**Student Number: 1802697**

**Abstract:** Nowadays we create a lot of Name Entity Recognition systems based on the Wikipedia data, but it depends on the enormous annotation works which is the bottleneck of our daily work. So, we decide to classify every Wikipedia article into name entity recognition format, then use the machine technique to train the model and get the accuracy.

Content

[1. Introduction 3](#_Toc1908259)

[2. Process 3](#_Toc1908260)

[**2.1.** **Re-process the Wikipedia article** 3](#_Toc1908261)

[**2.2.** **Training the Model** 4](#_Toc1908262)

[**2.3.** **Get the Results** 4](#_Toc1908263)

[3. Distant Learning 4](#_Toc1908264)

[**3.1.** **Definition** 4](#_Toc1908265)

[**3.2.** **Advantages** 4](#_Toc1908266)

[**3.3.** **Disadvantages** 5](#_Toc1908267)

[4. Approaches of NER 5](#_Toc1908268)

[4.1. Hand-made-rule-based NER 5](#_Toc1908269)

[4.2. Machine Learning Model 6](#_Toc1908270)

[ Supervised Learning 6](#_Toc1908271)

[ Un-supervised Learning 6](#_Toc1908272)

[4.3 Hybrid NER 6](#_Toc1908273)

[5. Conclusion 7](#_Toc1908274)

1. **Introduction**

Name Entity Recognition is an extraction of sentence that can identify the organization, person, time automatically. In the previous work we identify the name entity recognition by annotate the word from the Wikipedia article, there are enormous works that spend a lot of people and time. We decide to develop a NER method to identify the NER automatically. There are plenty of model to implement this task like perceptron, CRF++, and LibSVM, [1]because in our task we only need the words and the IOB format. So, I decide to use the perceptron to predict the NER.

1. **Process**
   1. **Re-process the Wikipedia article**

In our case we are going to use the Wikipedia article as our training data and testing data to build a NER system. First, we need to get the words and the IOB format of the article, after that we can put in the model to the training and testing. In this case we abstract the text from the aij-wikiner-en-wp2(A file that already collect the data from the Wikipedia). Then remove the pos\_tag of the data because the gold standard only needs the words and the IOB format.

* 1. **Training the Model**

In this assignment, I suppose to use the perceptron as the NER model. Perceptron is a linear model that use vector as the input and get the instance example as the output. Actually, the perceptron is a mapping function from the input space to the output space.

* 1. **Get the Results**

After we training the data, we can get the accuracy by using wikigold.conll.txt which is a gold standard dataset to test the model.

1. **Distant Learning** 
   1. **Definition**

Distant Learning is a learning model that abstract the message from the Wikipedia and training the model. It includes the positive data that from the mention-to-link Wikipedia pages and the negative data from the similar pages. Then train model by using the positive and negative example.

* 1. **Advantages**

We can easily find that the distant learning obtains lots of training data include the positive and negative example. By using these data, we can build a higher accuracy model. The prediction also can be close to the reality.

* 1. **Disadvantages**

Distant learning needs to find all related pages of one specific message, first of all, it needs lot of time than the other method. Secondly, the unfunctional requirement is higher than the other methods. Sometimes the distant learning also creates the overfitting problem due to it collect all the information from the Wikipedia.

1. **Approaches of NER**

As I said there are lot of methods applied to the NER. There are three main methods in identify the NER.[2] Hand-made-rule-based NER, Machine Learning NER and Hybrid NER.

## 4.1. Hand-made-rule-based NER

The hand-made-rule-based NER is extracting the information by using lots of human grammatically rules to identify the names, organization and locations. This method divide the process into three steps: Recognizing phrases, Recognizing patterns and Merging incidents[2]. This type can perform well in domain area and can predict the complex model, but it also has disadvantages. The model lacks the portability and if the data change slightly, the results can be catastrophic.

## 4.2. Machine Learning Model

There are two methods in Machine Learning Model: supervised learning and unsupervised learning.

* Supervised Learning

It is going to use the machine learning method to classify different label of data using same number of features. So, we have to clean the data first and give each data a label.

* Un-supervised Learning

In this method, the program is going to build the representation of the words. Then the representation is going to be used in compression, classify and so on.

## 4.3 Hybrid NER

The Hybrid NER looks like a combination of rule-base model and machine learning model. Then using strongest point to generate a new methods. Although this method have a better results than the other methods, it remain the weakness of the hand-made-rule-based NER

1. **Conclusion**

In this report, I discuss the advantages and disadvantages of the distant learning that applied to the NER. I also introduced some approaches of the name entity recognition. It still remain some function need to be improved in the future.

1. **Reference**

[1] J. Nothman, N. Ringland, W. Radford, T. Murphy, and J. R. Curran, “Learning multilingual named entity recognition from Wikipedia,” *Artif. Intell.*, vol. 194, pp. 151–175, Jan. 2013.

[2] A. Mansouri and L. Affendey, “Named entity recognition approaches,” *Ijcsns*, vol. 8, no. 2, pp. 339–344, 2008.