**A Survey of NER**

Xinyu Ma, Yifei Liu

Institute of Computing Technology

Chinese Academy of Science

Beijing, P.R.C

xinyuma2016@gmail.com

## Problem Definition

Named Entity Recognition is the problem of finding the members of various predetermined classes, such as person, organization, location, date/time, quantities, numbers etc. in text. [2]

It can be divided into two sub tasks: named entity identification (NEI) and named entity classification (NEC).

The NER project was first arrived in the sixth message understanding conference (MUC- 6) [3] and commit to identification of proper nouns (people and organizations), place names, temporal expressions and numerical expressions [4].

## Evaluation Metric

Different evaluation measures for examining the performance of NERC systems are discussed in the literature. The evaluation is done basically to check the ability of tool on finding correct entity types and their boundaries. For this, the NERC system’s predictions are compared with the predictions made by human annotators. The intrinsic evaluation metrics used for the comparison are Precision, Recall and F-score.

1. Precision, recall and F-score

Precision, Recall and F-score are calculated based on true positives (TP), false positives (FP) and false negatives (FN). True positives are the correctly labeled instances. False positives are the incorrectly labeled instances and false negatives are the missed-out instances by the system. F-score is the weighted mean of Precision and Recall. These metrics are formulated as given below.

In simple words, Precision is the ratio of correctly classified entities over total detected NEs. Recall is the ratio of relevant NEs over total detected entities by the system.

1. Matching predictions against Gold standard

Named Entity Detection involves finding the correct entity boundaries as well as a correct entity type. Most of the systems require an exact match on both entity type and boundary. The shared task for CoNLL-2003 [5] is one of the examples of exact matching.

The MUC-6 [3] events have defined a more loosened scheme which allows a partial credit for the systems finding correct boundaries regardless of the type as well as finding correct type regardless of boundaries using micro-averaged F-score.

ACE [6] has proposed the most complex form of evaluation, which is because each NE class is assigned a weighted parameter that contributes up to a maximum proportion of the final score. This measurement resolves the issues like a partial match, wrong type, etc. as well as considers the sub-types of NEs.

1. Macro-and micro averaged F-score

As most of the NERC systems involve multiple entities types, so it is often required to assess the performance of the system for all entity classes. Two measures are considered for this: Macro-averaged F-score and Micro-averaged F-score. Macro-averaged F-score is the average of F-scores of all entity classes in the corpus while micro-averaged F-score is calculated by adding the number of labeled entities together and then calculating Precision, Recall and F-score. The difference is that the micro-averaged measure can be badly affected by the larger classes in the corpus suppressing the performance of the system on smaller classes. However, MUC’s final score is calculated using micro-averaged F-score.

1. Cross-validation

Cross-validation is the measure used by different researchers [7,8,9]. This technique is the balanced version of evaluation, normally used for supervised learning methods. Cross-validation is based on the idea of dividing the dataset into n chunks and treating all the chunks except one for modeling the system. This process is usually repeated for k iterations that is why it is known as k-fold cross-validation. In each iteration, a different chunk is left and used for testing. The final score is calculated by averaging the results obtained in all iterations. Usually, 10-fold cross-validation is widely used for NERC task.

## Related Work

Earlier systems are most often based on hand-crafted rules as noted by [10]. These systems include usage of information lists such as gazetteers as well as rules based on syntactic-lexical patterns to identify and classify named entities.

Rule-based NERC systems are considered as highly efficient because they exploit the properties of language-related knowledge [11]. They employ domain specific features to obtain the enough accuracy. However, some limitations of these systems are that they are quite expensive, domain-specific and non-portable. Furthermore, these systems require human expertise regarding knowledge of the domain and language along with programming skills [12] for its development. Besides it, rule-based systems cannot be transferred across domains. Therefore, such kind of systems made for one domain cannot be ported into other domains which shifts the interests of researchers towards machine learning based approaches.

#### Learning-based approaches

Machine learning refers to the science of automatically learning complex patterns or sequence tagging algorithms which further makes efficient decisions about the data. Learning-based approaches can be divided into following three categories:

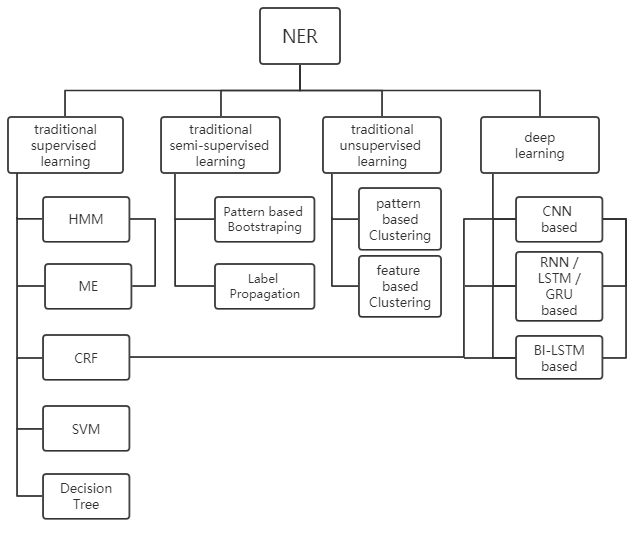


Figure 1.

1. Supervised learning

Supervised learning-based approaches are based on the idea of providing labeled training data involving positive and negative examples; constructing the adaptive features associative with examples; selecting appropriate learning algorithm that distinguishes positive from negative examples by consuming these features and recognizing similar information from unseen data.

Appropriate selection of features is a crucial task in supervised learning-based NERC systems. Features are the properties and attributes of textual objects in a computational model. Features play an important role to represent a multidimensional aspect of text forms which are further used by the learning methods for generating a model. This model can recognize patterns that find similar data and classifies positive and negative examples. Features in NERC task are well explained by [10]. The authors have classified the feature space into three groups namely list lookup features, document and corpus features and word-based features. List look-up features are based on linguistic resources such as lexicons, dictionaries, gazetteers, etc. These features determine whether a word is a member of any of these resources or not. Document and corpus related features are designed based on both document structure and content. Word-based features include orthographical, contextual and morphological features.

The choice of the learning algorithm is as well important as that of the feature selection. Various learning techniques have been used by different researchers for Named Entity Recognition Systems. Examples of these systems are: Hidden Markov Model (HMM) based systems [13], Support Vector Machine (SVM) based systems [14], Conditional Random Field (CRF) based systems [15,13], Maximum Entropy Markov Model (MEMM) based systems [16], Logistic Expression based systems [17], etc. Some classifier ensemble techniques [18] are also available in the literature.

1. Semi-supervised learning

Semi-supervised machine learning is a special form of learning. Traditional classifiers require a considerable amount of annotated training data. Annotation of training data is an expensive, difficult and time taking task because it requires the efforts of experienced human annotators. Semi-supervised learning addresses this problem by using both labeled and unlabeled corpus to make their hypothesis. These methods use a small number of training examples called ‘‘seed’’ for tagging unlabeled data. The results are then used to re-train the system to generate more labeled examples. This process continues to several times to make the learning decisions refined. The most popular method is ‘‘bootstrapping’’ [19] used by many researchers and gained popularity.

Semi-supervised pattern based bootstrapping approach has been used for identifying named entities from English and Tamil data [20]. In this approach, a small set of tagged training data is used to extract word and context features to define a five-word window context pattern for each named entity category. The identified patterns are used as seed patterns. These seed patterns are used to identify the entities as an exact match in the test set. Two parameters are used to decide the modification needed to generate new patterns. These parameters are the pattern scoring and the tuple value scoring. The pattern score determines which set of patterns are used for the next iteration. The tuple value scoring provides which set of tuple contributes to the named entity and decides the window movement that is a shift to the left or to the right and masks one tuple which generates new patterns that are used to learn new context to identify named entities.

1. Unsupervised learning

Unsupervised learning is an algorithm that uses information which is neither classified nor labeled. These methods purely use unlabeled data to make their decisions. The goal of unsupervised learning is to generate a model that considers the structural and distributional features of data to find more learning about the data.

The typical unsupervised approach is clustering and association rules-based approach. Clustering based approach uses distributional statistics to extract named entities out of unlabeled data by making use of context similarity. Association rules-based technique is concerned with finding associations amongst items within large databases. To deal with lack of annotated text across domains and languages, unsupervised techniques for NERC have been proposed [19].

#### Hybrid approaches

In the past few years, non-linear neural networks with as input distributed word representations, also known as word embeddings, have been broadly applied to NLP problems with great success. Collobert et al. (2011) proposed a simple but effective feed-forward neutral network that independently classifies labels for each word by using contexts within a window with fixed size. Recently, recurrent neural networks (RNN) (Goller and Kuchler, 1996), together with its variants such as long-short term memory (LSTM) (Hochreiter and Schmidhuber, 1997; Gers et al., 2000) and gated recurrent unit (GRU) (Cho et al., 2014), have shown great success in modeling sequential data. Several RNN-based neural network models have been proposed to solve sequence labeling tasks like speech recognition (Graves et al., 2013), POS tagging (Huang et al., 2015) and NER (Chiu and Nichols, 2015; Hu et al., 2016), achieving competitive performance against traditional models. However, even systems that have utilized distributed representations as inputs have used these to augment, rather than replace, hand-crafted features (e.g. word spelling and capitalization patterns). Their performance drops rapidly when the models solely depend on neural embeddings. An end-to-end architecture was proposed by XueZhe Ma et al.,2016 which is a combination of LSTM, CNN and CRF model.

Hybrid approaches hold the advantages of both learning-based and rule-based techniques. It finds the results by combining the results of two or more machine learning techniques or handcrafted rules. Various hybrid Named Entity Recognition Systems have been introduced by different researchers.

Hybrid systems are found more accurate than individual systems as observed in reviewing the literature.

## Benchmark

**Common dataset**: CoNLL-2003 dataset, MUC-7 dataset, KBP Track NER dataset.

Table 1. NER evaluation results of previous state-of-art method in CoNLL2003 dataset

|  |  |  |
| --- | --- | --- |
| author | model | F1 |
| Huang et al.(2015) | CRF | 83.02 |
| McCallum and Li(2003) | CRF+lexicons | 84.04 |
| Collobert et al.(2011) | CNN/NN+hand-drafted features | 89.59 |
| Huang et al.(2015) | BI-LSTM+CRF+ hand-drafted features | 90.10 |
| Chiu and Nichols(2015) | CNN(char+word)+BI-LSTM | 90.77 |
| Luo et.al(2015) | CRF+entity linking | 91.20 |
| Lample et.al(2016) | BI-LSTM(char+word)+CRF | 90.94 |
| Ma and Hovy(2016) | CNN(char+word)+BI-LSTM+CRF | 91.21 |
| Yang et.al(2017) | Transfer learning | 91.26 |
| Peters et al. (2018) | BiLSTM-CRF+ELMo | 92.22 |
| Akbik et al.(2018) | Contextual embedding+BI-LSTM +CRF(char + word ) | 93.09 |

Table 2. NER evaluation results of state-of-art NER systems in CoNLL2003 dataset

|  |  |  |  |
| --- | --- | --- | --- |
| System | Precision | Recall | F1 |
| Stanford | 95.1 | 78.3 | 85.9 |
| UIUC | 91.2 | 90.5 | 90.8 |
| Nerel | 86.8 | 89.5 | 88.2 |
| JERL | 91.5 | 91.4 | 91.2 |

The latest NLP-progress:

https://github.com/sebastianruder/NLP-progress/blob/master/named\_entity\_recognition.md

## Conclusion

From traditional machine learning method to deep learning method, we can see very explicit performance increase in CoNLL 2003 dataset. But most of the models today is just a combination of different neural network architectures. Some deep insight in this problem have not been found, there is still a long way to go.

## References

[1] Sakshi, Shailendra Singh, Dharam Vir, & Shashank Singh. "COMPARATIVE ANALYSIS OF VARIOUS APPROACHES BASED ON NAMED ENTITY RECOGNITION-A SURVEY." International Journal of Advanced Research in Computer Science [Online], 9.3 (2018): 132-138. Web. 4 Sep. 2018.

[2] Goyal, Archana, V. Gupta, and M. Kumar. "Recent Named Entity Recognition and Classification techniques: A systematic review." Computer Science Review 29(2018):21-43.

[3] R. Grishman, B. Sundheim, Message understanding conference-6: A brief history, in: COLING, 96, 1996, pp. 466–471.

[4] R.Sharnagat, "Named Entity Recognition: A Literature Survey", Center For Indian Language Technology, 2014.

[5] E.F.T.K. Sang, F.D. Meulder, Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition, in: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL, 4, Association for Computational Linguistics, 2003, pp. 142–147.

[6] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, R. Weischedel, The Automatic Content Extraction (ACE) program — tasks, data, and evaluation, in: LREC, 2, 2004, p. 1.

[7] X. Liu, M. Zhou, Two-stage NER for tweets with clustering, Inf. Process. Manag. 49 (1) (2013) 264–273.

[8] Y. Chen, T.A. Lasko, Q. Mei, J.C. Denny, H. Xu, A study of active learning methods for named entity recognition in clinical text, J. Biomed. Inform. 58 (2015) 11–18.

[9] M. Konkol, M. Konopík, Segment representations in named entity recognition, in: International Conference on Text, Speech, and Dialogue, Springer International Publishing, 2015, pp. 61–70.

[10] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Lingvist. Investig. 30 (1) (2007) 3–26.[11] K. Shaalan, Rule-based approach in Arabic natural language processing, Int. J. Inf. Commun. Technol. 3 (3) (2010) 11–19.

[12] S. Sarawagi, Information extraction, Found. Trends Databases 1 (3) (2008) 261–377.

[13] Y. Wang, Z. Yu, L. Chen, Y. Chen, Y. Liu, X. Hu, Y. Jiang, Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: An empirical study, J. Biomed. Inform. 47 (2014) 91–104.

[14] S.K. Saha, S. Narayan, S. Sarkar, P. Mitra, A composite kernel for named entity recognition, Pattern Recognit. Lett. 31 (12) (2010) 1591–1597.

[15] M. Majumder, U. Barman, R. Prasad, K. Saurabh, S.K. Saha, A novel technique for name identification from homeopathy diagnosis discussion forum, Procedia Technol. 6 (2012) 379–386.

[16] S.K. Saha, S. Sarkar, P. Mitra, Feature selection techniques for maximum entropy based biomedical named entity recognition, J. Biomed. Inform. 42 (5) (2009) 905–911.

[17] T. Ek, C. Kirkegaard, H. Jonsson, P. Nugues, Named entity recognition for short text messages, Procedia Soc. Behav. Sci. 27 (2011) 178–187.[18] Z. Munkhjargal, G. Bella, A. Chagnaa, F. Giunchiglia, Named entity recognition for Mongolian language, in: International Conference on Text, Speech, and Dialogue, Springer International Publishing, 2015, pp. 243–251.

[19] M. Bhagavatula, S GSK, V. Varma, Named entity recognition an aid to improve multilingual entity filling in language-independent approach, in: Proceed ings of the First Workshop on Information and Knowledge Management for Developing Region, ACM, 2012, pp. 3–10.

[20] B. Jagan, S. Thenmalar, T.V. Geetha, Semi-supervised bootstrapping approach for named entity recognition, CoRR (2015). abs/1511.06833.

[21] L. Li, R. Zhou, D. Huang, Two-phase biomedical named entity recognition using CRFs, Comput. Biol. Chem. 33 (4) (2009) 334–338.

[22] Z. Guanming, Z. Chuang, X. Bo, L. Zhiqing, CRFS-based Chinese named entity recognition with improved tag set, in: 2009 WRI World Congress on Computer Science and Information Engineering, 2009.

[23] A. Ekbal, S. Saha, A multiobjective simulated annealing approach for classifier ensemble: Named entity recognition in Indian languages as case studies, Expert Syst. Appl. 38 (12) (2011) 14760–14772.

[24] D. Küçük, A. Yazıcı, A hybrid named entity recognizer for Turkish, Expert Syst. Appl. 39 (3) (2012) 2733–2742.

[25] X. Liu, M. Zhou, Two-stage NER for tweets with clustering, Inf. Process. Manag. 49 (1) (2013) 264–273.