Teaching Machines to Find Names

Raymond Chiong
Swinburne University of Technology, Sarawak Campus, Malaysia

INTRODUCTION

In the field of Natural Language Processing, one of the very important research areas of Information Extraction (IE) comes in Named Entity Recognition (NER). NER is a subtask of IE that seeks to identify and classify the predefined categories of named entities in text documents. Considerable amount of work has been done on NER in recent years due to the increasing demand of automated texts and the wide availability of electronic corpora. While it is relatively easy and natural for a human reader to read and understand the context of a given article, getting a machine to understand and differentiate between words is a big challenge. For instance, the word 'brown' may refer to a person called Mr. Brown, or the colour of an item which is brown. Human readers can easily discern the meaning of the word by looking at the context of that particular sentence, but it would be almost impossible for a computer to interpret it without any additional information.

To deal with the issue, researchers in NER field have proposed various rule-based systems (Wakao, Gaizauskas & Wilks, 1996; Krupka & Hausman, 1998; Maynard, Tablan, Ursu, Cunningham & Wilks, 2001). These systems are able to achieve high accuracy in recognition with the help of some lists of known named entities called gazetteers. The problem with rule-based approach is that it lacks the robustness and portability. It incurs steep maintenance cost especially when new rules need to be introduced for some new information or new domains.

A better option is thus to use machine learning approach that is trainable and adaptable. Three well-known machine learning approaches that have been used extensively in NER are Hidden Markov Model (HMM), Maximum Entropy Model (MEM) and Decision Tree. Many of the existing machine learning-based NER systems (Bikel, Schwartz & Weischedel, 1999; Zhou & Su, 2002; Borthwick, Sterling, Agichten & Grisham, 1998; Bender, Och & Ney, 2003; Chieu & Ng, 2002; Sekine, Grisham & Shinnou, 1998) are able to achieve near-human performance for named entity tagging, even though the overall performance is still about 2% short from the rule-based systems.

There have also been many attempts to improve the performance of NER using a hybrid approach with the combination of handcrafted rules and statistical models (Mikheev, Moens & Grover, 1999; Srihari & Li, 2000; Seon, Ko, Kim & Seo, 2001). These systems can achieve relatively good performance in the targeted domains owing to the comprehensive handcrafted rules. Nevertheless, the portability problem still remains unsolved when it comes to dealing with NER in various domains.

As such, this article presents a hybrid machine learning approach using MEM and HMM successively. The reason for using two statistical models in succession instead of one is due to the distinctive nature of the two models. HMM is able to achieve better performance than any other statistical models, and is generally regarded as the most successful one in machine learning approach. However, it suffers from sparseness problem, which means considerable amount of data is needed for it to achieve acceptable performance. On the other hand, MEM is able to maintain reasonable performance even when there is little data available for training purpose. The idea is therefore to walkthrough the testing corpus using MEM first in order to generate a temporary tagging result, while this procedure can be simultaneously used as a training process for HMM. During the second walkthrough, the corpus uses HMM for the final tagging. In this process, the temporary tagging result generated by MEM will be used as a reference for subsequent error checking and correction. In the case when there is little training data available, the final result can still be reliable based on the contribution of the initial MEM tagging result.

BACKGROUND

Message Understanding Conference

In 1987, the Naval Ocean Systems Center (NOSC), which is presently known as the Naval Command,

Control and Ocean Surveillance Center, initiated the first Message Understanding Conference (MUC). Subsequently, a series of MUCs had been held and designed to promote and evaluate research in IE. The evaluations achieved through these MUCs have led the research program in IE until its present state.

In 1995, goals and tasks were set up for MUC-6 to make the IE system more practical with an aim to achieve automatic performance with high accuracy. "Named Entity" was then developed to help identifying the names of persons, organizations, and geographic locations in a text. Since then, the NER tasks have become a central theme in MUC (see Chinchor, 1995 and Chinchor, 1998 for more details).

According to the specifications defined by MUC, the NER tasks generally work on seven types of named entities as listed below with their respective markup:

- PERSON (ENAMEX)
- ORGANIZATION (ENAMEX)
- LOCATION (ENAMEX)
- DATE (TIMEX)
- TIME (TIMEX)
- MONEY (NUMEX)
- PERCENT (NUMEX)

From the list above, three subtasks are derived from these seven types of named entities and assigned with three respective SGML tag elements, namely ENAMEX, TIMEX and NUMEX. As TIMEX and NUMEX are fairly easy to predict with some effective finite state methods (Roche & Schabes, 1997), most of the current research deals only with ENAMEX which are highly variable and ambiguous.

Previous Approaches

Since MUC-6 and MUC-7, many NER systems have been proposed and proven to be successful in their targeted domains. In general, NER systems that use handcrafted rules still lead the way, with the highest F-measure score up to 96.4% achieved in MUC-6 as compared to the statistical approaches that were able to achieve 94.9% (Zhou & Su, 2002).

In rule-based approach, a set of rules or patterns is defined to identify the named entities in a text. These rules or patterns consist of distinctive word format, such as capitalization or particular preposition prior to a named entity. For instance, a capitalized string behind titles such as 'Mr', 'Dr', etc will be identified as name of a person, whereas a capitalized word after a preposition such as 'in', 'at', 'near', etc is most likely to be a location. By implementing a finite set of carefully predefined pattern matching rules, the named entities within a text could be found systematically.

There have been substantial amount of works done using the rule-based approach. One of the very well documented systems that followed the direction of this approach was the framework of the LaSIE System reported by Wakao et al. (1996). Another well-known example of rule-based system can be found in the IsoQuest's NetOwl Text Extraction System presented by Krupka and Hausman (1998). Meanwhile, Diana Maynard et al. (2001) had also built an NER system based on handcrafted rules that is able to achieve an average of 93% precision and 95% recall across diverse text types.

Statistical approach, on the other hand, works by using a probabilistic model containing features to the data which are similar to the rule-based approach. The features of the data, which could be understood as rules set for the probabilistic model, are produced by learning the resulting corpora with correctly marked named entities. The probabilistic model then uses the features to calculate and identify the most probable named entities. As such, if the annotated features of the data are truly reliable, the model would have a high probability in finding almost all the named entities within a text.

In the last decade, large amount of works in NER have been done using the statistical approach based on some very large corpora. The MEM, one of the most popular statistical models, has been applied frequently in various NER tasks. One significant account on MEM is the MENE system reported by Borthwick et al. (1998). In their system, they used four main features to identify the named entities, which they referred to as the binary features, lexical features, section features and dictionary features.

The binary features in MENE system basically deal with capitalization in the text. Meanwhile, lexical features are concerned with the lexical terms such as list of words and their types which are used with a grammar. Section features indicate a current section of the text, whereas the dictionary features make use of a broad array of dictionaries of single or multiple terms such as first names, organization names, corporate suffixes, etc. The dictionary features are similar to the gazetteers