Person-specific named entity recognition using SVM with rich feature sets*

Hui NIE†

School of Information Management, Sun Yat-sen University, Guangzhou 510275, China

Abstract

Purpose: The purpose of the study is to explore the potential use of nature language process (NLP) and machine learning (ML) techniques and intents to find a feasible strategy and effective approach to fulfill the NER task for Web oriented person-specific information extraction.

Design/methodology/approach: An SVM-based multi-classification approach combined with a set of rich NLP features derived from state-of-the-art NLP techniques has been proposed to fulfill the NER task. A group of experiments has been designed to investigate the influence of various NLP-based features to the performance of the system, especially the semantic features. Optimal parameter settings regarding with SVM models, including kernel functions, margin parameter of SVM model and the context window size, have been explored through experiments as well.

Findings: The SVM-based multi-classification approach has been proved to be effective for the NER task. This work shows that NLP-based features are of great importance in data-driven NE recognition, particularly the semantic features. The study indicates that higher order kernel function may not be desirable for the specific classification problem in practical application. The simple linear-kernel SVM model performed better in this case. Moreover, the modified SVM models with uneven margin parameter are more common and flexible, which have been proved to solve the imbalanced data problem better.

Research limitations/implications: The SVM-based approach for NER problem is only proved to be effective on limited experiment data. Further research need to be conducted on the large batch of real Web data. In addition, the performance of the NER system need be tested when incorporated into a complete IE framework.

Originality/value: The specially designed experiments make it feasible to fully explore the characters of the data and obtain the optimal parameter settings for the NER task, leading to a preferable rate in recall, precision and $F_1$ measures. The overall system performance ($F_1$ value) for all types of name entities can achieve above 88.6%, which can meet the requirements for the practical application.

* This work is support by the Special Research Fundation for Young Teachers of Sun Yat-sen University (Grant No. 2000-3161101) and Humanity and Social Science Youth Foundation of Ministry of Education of China (Grant No. 08JC870013).
† Correspondence: Hui Nie (E-mail: issnh@mail.sysu.edu.cn).
Keywords  Named entity recognition; Natural language processing; SVM-based classifier; Feature selection

1 Introduction

Information extraction is an important task in natural language processing (NLP). It is aimed at scanning text for information of interest, including entities and the relations among them. Information extraction has many practical applications. With the emergence of SSN analysis and WePS application, the research on person-specific information extraction have become active due to the rapid growth of Web resources. The goal of SSN analysis is to provide deeper insight into the personal connections in academic field by analyzing relations among researchers[1,2]. Actually, a large amount of personal information can be collected from Internet. With regard to WePS, the new Web service has come to realize the discrimination of Web people by identifying the relevant personal information.

Although the purpose of applications is technically different, the major task of SSN and WePS is the detection and recognition of Web-based personal information, which is normally carried out by a person-specific information extraction system, whose performance mainly relies on such techniques as NER, text mining, pattern matching, and relation discovery. Among these, NER plays an important role in the whole information searching process. As described by Paul Kalmar[3], when dealing with an unstructured text, an NER system is critical to provide inputs for entity disambiguation, and semantic related features play the significant role in entity definition and extraction. Given the practical application of personal information collection for talent people, in this article, great emphases are particularly placed on the investigation of NER techniques in the person-specific information extraction task.

2 Related research

2.1 NER

NER techniques can be classified into two categories: Knowledge-based approaches and machine learning approaches. The former normally makes use of human knowledge in developing grammar rules and dictionaries, which may require no or very little training data, and is time-consuming and expensive[4]; while the latter aims at reducing human effort in maintaining rules and dictionaries. Due to such features as being easily trainable, adoptable to different domains and languages as well as less expensive for maintenance, they are widely used in NER nowadays[5].

In the machine learning area, NER is typically viewed as a sequence-labeling problem. Its common solution is to use machine learning with point-wise
Person-specific named entity recognition using SVM with rich feature sets

Hui NIE

Research Paper

National Science Library,
Chinese Academy of Sciences
http://www.chinalibraries.net

classification, which breaks the sequence into a number of data points and solves the point-wise classification problems independently. The normally used techniques are support vector machine (SVM), maximum entropy, naïve Bayes, k-nearest neighbour and conditional random fields (CRFs). Among them, SVM is one of the most successful approaches, which has superior performance on many NLP-based classification tasks[6,7]. For instance, Isozaki[6] compared the three commonly used methods, SVM, maximal entropy and rule-based learning systems, for NER he found that the SVM-based system outperformed the other two.

SVMs are well-known for their good generalization performance and SVM-based classification is often reported to have achieved high accuracy without falling into over-fitting even though with a large number of words taken as the features[8]. The first SVM-based NER system proposed by Yamada et al.[9] for Japanese gave the best performance at CoNLL-2000 shared tasks. Due to the ability in dealing with the diverse and overlapping features of natural languages, basically, SVMs have been one of the current dominant techniques for addressing NER. The lastest SVM-based NER related study can be found in Refs. [8] and [10–12].

Asif et al.[8] presented an SVM-based NER system for specific Indian languages, which fully took advantage of the power of SVMs for handling various language features. The best set of features for NER in Indian language has been found out and evaluation results demonstrated the overall $F_1$ values of 84.15%. Studies of Ju et al.[10] and Doan et al.[11] focused on NER in biomedical field. An SVM-based model has been developed to identify biomedical entities, and various types of features for NER in clinical text have been systematically investigated[11]. Doan et al.[11] particularly highlighted the importance of semantic features in NER. Experiment results demonstrated that the SVM-based NER system can achieve the best $F_1$ (90.05%) when semantic features were included. While in the work of Saha et al.[12], a novel kernel function for SVM was specifically designed for string based features. The results are quite promising in both general language and biomedical NER tasks. From these representative work, it is clear that the semantic features play a critical role in SVM-based models.

Actually, the quality of NER models, to a large extent, depends on the power of the data representation; that is, the features. Features selection has always been the most popular technical issue in ML-based NER task.

The generally used features are classified into three types: Lexical features, syntactic features and semantic features. The lexical feature type mainly considers the word appearance in the text, such as word string and word morphological patterns, which is the most common feature type used in NER. Syntactic features represent the structural properties of the text (e.g. post-of-speech tag) beyond the lexical level, which are more flexible and richer. With regard to the semantic
features, many studies[11-13] showed that they had highly positive impacts on performance of NER system because they consider the meaning of words. The work relevant to features selection can be found in the research of Refs. [3, 7] and [15, 16].

2.2 NER in person-specific IE

From the application point of view, the person-specific IE research can be reviewed by exploring the studies about SSN and WebPS due to the most immediate application relations between them.

SSN related applications intend to mining person-specific entities and relations between them[17]. Since the person-specific data come from huge amount of Web pages, SSN projects commonly adopt lightly-supervised methods. By reviewing SSN related extraction systems, it has been found that the input of the systems is normally a seed list of target information pairs. The pairs may contain related person-specific entities (e.g. person and publication)[18] or entity-attribute pairs (e.g. person-publication). The extraction rules for entities or their relations are learned by an iterative learning processing. The self-training based IE systems assume that important information can be found at a number of places and in various forms on the Web, therefore, multiple accurate extraction rules can be obtained to collect the specific entity information[19].

For person-specific information extraction in WePS[20-22], the most relevant study is called attributes extraction (AE) task. With regard to the task, a two-step approach was proposed by Watanabe[23], in which the potential attributes strings in a given text were annotated firstly, then the attribute values relevant to the target person were selected. Technically, various available NER tools[18] have been utilized to identify person-specific entities and different attributes have been handled by various approaches. For example, OpenNLP\(^\text{a}\) combined with regular expression (RegEx) and Gazetteer-based matching can be used to recognize and extract such entity attributes as occupation, degree, location and organization, etc[24,25].

3 Research gaps and objective

The above relevant WePS and SSN studies show that a person-specific NER has always been the most important part for identifying attributes of individuals and detecting relationships among individual-specific entities. Based on the different application purposes, both research applications adopted different techniques to address their person-specific NER problems. Studies of SSN focused on solving how to precisely extract person-specific information from a large amount Web

---

resource efficiently and how the lightly-supervised ML methods have been employed. Since studies regarding WePS could highlight the discrimination of Web people, most studies made use of the existing NER tools to carry out the people clustering task. Personal information entities identification techniques have not been particularly investigated.

In the view point of application, this research is under the topic of SSN analysis. The research is inspired by the efficient learning methods of SSN, but is applied in different scenarios. In this paper, information for a specific individual is supposed to be available mainly from individual homepages. The research target is to capture as much information about the given talent people as possible from the limited Web page resources. Lightly-supervised approaches might be able to achieve high precision in the case but relatively low recall information extraction due to the specific resources. Actually, optimal recall is targeted as well. Additionally, even a bunch of existing research on NER techniques has been conducted for many years, most researches were generally based on some available benchmarks, e.g. MUC dataset for general language and i2b2\textsuperscript{2} datasets for biomedical NER tasks, the use of ML to NER for data resources collected from Web is quite new.

In view of the research gaps, the person-specific IE approach study has been conducted. The three main research questions are posed as follows:

*Question 1:* For given Web-based resources, what is an appropriate data-processing strategy for person-specific NER by using SVM-based classification techniques?

*Question 2:* What kind of NLP-based features and ways of combination might be most helpful for improving the performance of the NER system?

*Question 3:* With regard to the SVM model, what is the impact exerted by the important parameters on the performance of system and how could we obtain the optimal parameter settings by experiments?

This research is focused on exploring the potential use of NLP features and ML techniques and intents to find a feasible strategy to tackle the task of Web oriented information extraction for talent people. A group of experiments has been designed to investigate the influence of various NLP-based features as well as important parameters. The performance of the system and that with different feature sets and classifier model parameters have been tested and evaluated by typical precision, recall and $F_1$ values. Better performance (e.g. $F_1 \geq 80\%$) was expected to obtain for practical application requirement.

4 Research design

To address questions above, an SVM-based learning system has been designed to recognize person-specific entities from Web pages. Specifically, experiments have been conducted on the system for exploring the performance changes when different features were added one by one and combined with different SVM models. Fig.1 depicts the system design, where the NER task is carried out in four stages, i.e., feature collection, semantic annotation, learning & recognition and evaluation.

In the stage of feature collection, a set of NLP-based plugins provided by GATE® has been fully utilized to extract various language features according to the data processing flow proposed in Fig.1. Two crucial annotation components have been implemented in the second stage by means of lexicons and rule-based annotation engine. A set of person-specific words and phrases has been extracted from the personal Web pages by using regular expression to create a group of Gazetteers, the list-based annotation lexicons. Meanwhile, rule-based approaches were further combined to generate the semantic labels. Though lack of existing lexicons, semantic features have been obtained with less manual work, which would exert significant influence on the performance improvement of the system.

Fig. 1 The framework of the IE system.

In the third stage, a multi-classification SVM model has been proposed to address the sequence-labeling problem. Given the imbalanced dataset, a parameter for uneven margins $\tau$ was introduced into the standard model, which made the modified SVM model more general and flexible. Moreover, different SVM kernel functions have been investigated. Specific experiments are designed to find the optimal SVM model for the NER task.

4.1 Pre-processing

Normally, information on webpage is presented in semi-structured or raw text forms. The first stage of the work is to filter out the noise information from the HTML-style webpage and to extract the major contents in textual form as the analysis objects. This can be achieved by using webpage segmentation and topical detection techniques. Although the techniques will not be discussed in detail in this paper, textual Web content extraction is always the necessary pre-processing stage for entity recognition task.

4.2 Feature collection

During feature collection, NLP tools[26], which are called processing resources (PRs) in GATE, have been employed to perform the fundamental text processing as well as to derive NL features for SVM model training.

- **Document reset PR** enables the document to be reset to its original state.
- **Sentence splitter** is a cascade of multiple finite-state transducers which can segment the text into sentences. This module is required for the tagger.
- **Tokeniser** splits the text into very simple tokens such as numbers, punctuations and words, which are the basic units for the analysis that follows.
- **OrthoMatcher** is used to add identity relations between the named entities found by the semantic tags, so that co-reference can be performed.
- **POSTagger** produces a part-of-speech tag as an annotation on each word. For instance, a word tagged with “NP” indicates the word is a singular proper noun.
- **Morphological analyser** is used to identify a word’s lemma and affix. These values are included as additional features, such as the “root” on the token annotation[26].

Using the above plugin-based NLP tools of GATE, most NL features required for generating the feature vectors can be derived, such as token kind, string, POS, length, root, and orth (indicating upper or lower case of words).
4.3 Semantic annotation

During the semantic annotation, additional semantic features would have been added by utilizing Gazetteers. In GATE, Gazetteers are the set of lists containing names of entities which are used to find occurrences of names in the text, commonly for task of NER\(^2\). Taking the research for example, a set of Gazetteers especially for the four target entity types, i.e., person, organization, title and degree, has been built firstly. In target texts, each entity word appeared in the Gazetteers would have been labeled “Lookup”. The “majorType” attribute of “Lookup” indicates the entity type of the word being referred to. The dictionary-based entity detection mechanism of GATE enables the semantic features of words being annotated easily. Peculiarly, some general entity types such as “date” could be automatically labeled by using the rule-based annotation engine. Finally, the results of features extraction can be observed directly in the annotation window, as shown in Fig. 2.

![Fig. 2 Features extraction results.](image)

From Fig. 2, the extracted features for the example “Associate” can be seen as follows:

- category = NNP, kind = word, length = 9, orth = upper-initial
- root = associate, string = Associate

Another example of the word “Professor” has been recognized as a “Lookup” according to the related Gazetteer, and the features are:

- majorType = title, minorType = academic

4.4 Learning & recognition

Once all features have been obtained, the training feature vectors could be fed into the learning system. The result of an NER example is shown in Fig. 3.
Person-specific named entity recognition using SVM with rich feature sets

5 Methodology

In this paper, an SVM-based algorithm is adopted to extract person-specific information. The information extraction task is converted to a multi-classification problem.

5.1 Support vector machine

SVM is a general supervised learning algorithm which achieves state-of-the-art results for many classification tasks including NER. The algorithm description presented here is based on the work of Hearst[27], the most representative article about SVM.

Suppose we have a set of training data for a two-class problem: \((x_1, y_1), \ldots, (x_n, y_n)\), where \(x_i\) is a feature vector of the \(i\)th sample in the training data and \(y_i \in \{+1, -1\}\) is the label for the sample. The goal is to find a decision function that accurately predicts \(y\) for unseen \(x\). An SVM classifier gives a decision function \(f(x)\) for an input vector \(x\), where

\[
y = f(x) = \text{sign}(g(x))
\]

\[
g(x) = \sum_{i=1}^{n} v_i \cdot k(x, z_i) + b
\]

The parameter \(v_i\) is computed as the solution of a quadratic programming problem[27]. Here, \(f(x) = +1\), which means \(x\) is a member of a certain class; and \(f(x) = -1\) means...

Figure 3 shows that the strings labeled with “Entity” are the target entities identified by the classification model. For example, “University of California” has been classified as an organization entity, which has been indicated by the value of the attribution “class”.

Fig. 3 An example of NER.
$x$ is not a member. $z_i (i=1,...,l)$ are called support vectors and representatives of training examples. $k(x, z)$ is a kernel function that implicitly maps vectors into a higher dimensional space. Typical kernels use dot products, i.e., $k(x, z) = (x \cdot z)$. A polynomial kernel of degree $d$ is given by $k(x, z) = (x \cdot z)^d$. Selection of the optimal kernel depends on the particular application. In our experiments, linear kernel and quadratic kernel have been tested and compared. The results showed that in spite of simplicity, linear kernel achieved better performance than quadratic kernel in the NER task.

In addition, inspired by the related work of Li\cite{13}, an SVM model with uneven margins has been adopted due to its better performance compared to the standard SVM model on imbalanced dataset. Specifically, the standard SVM treats positive and negative examples equally so that the margin of the SVM hyperplane to the negative training examples is equal to the margin to the positive training examples.

However, for imbalanced training data, where the positive examples are so rare that they may not represent the genuine distribution of the positive examples, a larger positive margin rather than the negative would be beneficial. Hence, it is preferable to generalize the standard SVM by introducing an uneven margin parameter, which is the ratio of the negative margin to the positive margin. The uneven margin parameter is 1 for the standard SVM and 0~1 for uneven margin SVM. The value can be empirically determined, for example, using $n$-fold cross-validation on training set or hold-out development set. Given our NER task usually handles limited data sets and thus often imbalanced training data, we expect to benefit more from the uneven margin SVMs than the standard SVMs.

### 5.2 Multi-classification SVM model for NER

Typical SVM models are binary classifiers. However, named entities in the task usually span multiple words. Therefore, more than one classifier need to be trained to identify the whole entity chunk, which is the common approach for SVM-based multi-classification task. Taking the NER system in the paper for example, three SVM classifiers have been trained to identify each entity type. One classifier is for identifying the start word of the entity, one for the end word, and the third for the single word entity. The start, end and single words of an entity type are called “entity sign”. For instance, if we want to identify the “University of Washington” as an “Organization” entity, the words in the chunk are classified as follows: “University” = Org-Start, “Washington” = Org-End. The word “Oxford” is recognized as a single-word “Organization” entity, therefore, the “entity sign” of “Oxford” is Org-single. Thus, since each entity type needs three classifiers, the NER task, which requires 4 types of entities (person, organization, title, degree) to be extracted, would need 3*4=12 classifiers in total.
As general SVMs can only solve binary classification problems, a “one class versus others” strategy has been employed to address the multi-classification problem in the task. That is, each classifier $f_c(x)$ is trained to distinguish the members of a class “$c$” from the non-members. Thus, an unseen vector might be labeled belonging to multi-categories. To avoid such multi-labeling, one common way is to compare $g_c(x)$ values and choose the class index “$c$” corresponding to the largest $g_c(x)$. This can be carried out by using a sigmoid function to transfer $g_c(x)$ into a probability, then applying the viterbi algorithm to determine the optimal class label for the word[6].

5.3 Feature selection

Another issue which is as important as classification model is feature selection. In NER task, each word in the document is regarded as an individual instance and classified (e.g., entity types). Regarding SVM-based classifiers, words are represented by feature vectors no matter whether they are training samples or test samples. Thus, selecting the features which best characterize words is critical to achieve optimal classifier performance. Existing researches[7] suggested that linguistic features of words play the most important role in NER task.

The commonly used features are classified into three types: Lexical features, syntactic features and semantic features. Lexical features describe the word appearance in the text. Four NLP features - token forms, capitalization information, token kind and lemma of words were considered in our research. Syntactic features represent structural properties of the text, e.g., POS tag. As for semantic information, we particularly considered the certain predefined semantic classes of the words, such as “Lookup” feature derived from predefined Gazetteers. All of the features above were obtained by fully utilizing the corresponding GATE plugins. Table 1 shows a sample text with its associated features defined in the GATE.

<table>
<thead>
<tr>
<th>String</th>
<th>Case**</th>
<th>Tokenkind***</th>
<th>Length of token</th>
<th>POS</th>
<th>Lookup†</th>
<th>Position++</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ph D. (Harvard)</td>
<td>Upper-initial</td>
<td>Word</td>
<td>2</td>
<td>NNP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>Punctuation</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.</td>
<td>Word</td>
<td>1</td>
<td>NNP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(</td>
<td>Punctuation</td>
<td>1</td>
<td>Startpunct</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>)</td>
<td>Word</td>
<td>7</td>
<td>NNP</td>
<td>Organization</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Punctuation</td>
<td>1</td>
<td>Endpunct</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: *, Token forms; **, orth; ***, simple token kinds; †, semantic types from Gazetteer lists; and ++, distinctive feature for bracket typed signs.
From Table 1, we can observe that the string “Ph.D” cannot be recognized as a
degree-related term since it has not been listed in the relevant GATE Gazetteers. If
the relevant Gazetteer can be extended, more semantic features will be available,
and the precision of NER will make improvements. Generally, there are two steps
for a feature vector to be derived from the NL features of each token[13].

- Step 1: All possible features from the training documents are collected and
each feature is indexed with a unique identifier. Each dimension of the feature
vector corresponds to one feature.
- Step-2: For each token, each component of the feature vector that corresponds
to the value of the respective NL feature is set to be 1, and all other components
are set to 0.

Regarding the free-text oriented information extraction, the context information
around tokens is usually as important as tokens itself. Therefore, in addition to the
features of the given token, features of the preceding and following tokens should
be included in the input feature vector as well. The parameter “context window size”
is used to set the number of neighboring tokens around the current token. Suppose
the window size is \( N \), the feature vector will be derived from \( 2*N+1 \) tokens, which
include the \( N \) preceding tokens, the current tokens, and the \( N \) following tokens. A
raw text is given as an example in the following:

“Professor Houk received his A.B., M.S., and PhD degrees at Harvard University,…”

The current token is “degrees” and the value of the “context window size” is 3. Then the feature vector of “degrees” will be combine the features from the 7 adjacent
tokens, as shown in Fig. 4.

5.4 Post-processing

As the three trained SVM classifiers only identify the “entity sign”, i.e., the start,
end or single words of a target entity chunk, post-processing is needed to identify
the whole entity form these sign words and assign a single label to it. Obviously,
the single word can be annotated directly according to its category identified by the
SVM classifier. While the start and end words need to be combined through a group
consistency inspection before a single label can be assigned. The inspection is
performed as described by Freitag et al. in the following two steps[28]:
Person-specific named entity recognition using SVM with rich feature sets

Hui NIE

Research Paper

National Science Library,
Chinese Academy of Sciences

- Firstly, the target text is scanned to remove the start labels without the matching end label, and the end labels without the preceding start label.
- Then the remaining candidate entities, which include the matching start and end words and the words in between, are filtered in terms of their lengths.

Specifically, the label of a candidate entity are removed if its length is not equal to that of any entity of the same type in the training data.

6 Experiments study

6.1 Test bed

In order to examine the SVM-based approach for personal information extraction, a group of experiments was conducted on a corpus of personal Web pages collected automatically from websites of 10 universities by using Heritrix®. By pre-processing, noisy and misleading elements, such as menu elements in the Web pages have been filtered out firstly, then 225 Web pages with complete description about specific persons have been selected as the test bed. Four MIS graduate students manually annotated the person-related entities with the types of person, title, degree and organization (this research did not take the available benchmarks, e.g. MUC data as the test bed due to the field-specific applications). Since the Web pages were extracted manually, good-quality experiment data could be ensured. Table 2 shows the statistics of our test bed.

<table>
<thead>
<tr>
<th>Item</th>
<th>Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Web pages</td>
<td>225</td>
</tr>
<tr>
<td>Number of labeled words</td>
<td>1,784</td>
</tr>
<tr>
<td>Number of words labeled with person</td>
<td>473</td>
</tr>
<tr>
<td>Number of words labeled with title</td>
<td>389</td>
</tr>
<tr>
<td>Number of words labeled with degree</td>
<td>247</td>
</tr>
<tr>
<td>Number of words labeled with organization</td>
<td>675</td>
</tr>
</tbody>
</table>

The experiments were conducted on the test bed. The performance of the whole system as well as that with different feature sets and classifier model parameters were evaluated by 10-fold cross-validation.

6.2 Experiment parameter settings

A series of preliminary tests were carried out to investigate the influence of different parameters and determine the optimal parameter settings for our system. The

---

standard performance measures are used, including recall, precision and \( F_1 \), where \( F_1 \) is the harmonic mean of precision and recall, i.e.,

\[
F_1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}.
\]

When calculating precision, a system-marked tag is considered correct if it matches exactly the manually labeled tag, both in terms of types and the start and end offsets in the context.

The parameters investigated are feature sets, SVM model settings and the context window size. As exhaustive testing of all parameters settings is impractical, these parameters were investigated sequentially according to their importance of the performance in our experiments.

6.2.1 Feature sets

Since feature selection is generally considered the most important issue for SVM-based classifiers, the impact of different features on the system performance has been focused on investigation firstly. The study started with a baseline experiment where “string” was used as the only feature and the system performance was evaluated. Then, different features were gradually added to the initial feature set and the system performance was monitored. In this group of experiments, standard linear SVM model (linear kernel, uneven margin \( \tau = 1.0 \)) and the default context window size = 5 were employed. Table 3 shows the system performance using the different feature sets.

<table>
<thead>
<tr>
<th>Feature sets</th>
<th>Precision</th>
<th>Recall</th>
<th>( F_1 )</th>
<th>Improvement of ( F_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>String</td>
<td>0.82592595</td>
<td>0.57343644</td>
<td>0.6768979</td>
<td>–</td>
</tr>
<tr>
<td>String+Orth</td>
<td>0.80980045</td>
<td>0.62833536</td>
<td>0.7032732</td>
<td>0.02637</td>
</tr>
<tr>
<td>String+Orth+Kind</td>
<td>0.80124104</td>
<td>0.62984645</td>
<td>0.7006065</td>
<td>–0.00267</td>
</tr>
<tr>
<td>String+Orth+Root</td>
<td>0.78856575</td>
<td>0.65920097</td>
<td>0.71334255</td>
<td>0.01007</td>
</tr>
<tr>
<td>String+Orth+Root+POS</td>
<td>0.79222655</td>
<td>0.60444164</td>
<td>0.685671</td>
<td>–0.02767</td>
</tr>
<tr>
<td>String+Orth+Root+Lookup</td>
<td>0.9611278</td>
<td>0.94789934</td>
<td>0.9544495</td>
<td>0.24111</td>
</tr>
</tbody>
</table>

Note: SVM linear kernel (\( \tau = 1.0 \)); and the context window size = 5.

From Table 3 we can see that the system performance has generally improved as more features have been added. Compared to the single feature “String”, the feature set “String+Orth+Root+Lookup” increases the \( F_1 \) measurement from 0.6768979 to 0.9544495. Particularly, after adding the semantic feature “Lookup”, \( F_1 \) is increased by 0.24111. This shows a significant improvement. Semantic features are particularly beneficial for NE extraction and worth paying more attention to. In contrast, it is also noted that the feature “Tokenkind” and “POS” are not as helpful as expected and result in the decreases of \( F_1 \). Therefore, they were eliminated from further
experiments. This suggests that some features may not have positive contribution to the NER tasks in some specific field.

### 6.2.2 SVM kernels & uneven margins parameter

After identifying the most effective feature set, further studies were performed to find the optimal SVM model parameters. The author focused on studying the kernel function and the uneven margins parameter ($\tau$) as they are the two key parameters which can fine tune the SVM model to suit the training data as represented by the selected feature set.

- **Kernel function.** The kernel function is generally associated with the complexity of the space spanned by the feature vectors of the training data. Usually, higher order kernel function indicates greater complexity of the space. Table 4 compared two basic kernel functions, i.e., the linear and quadratic functions.

- **Uneven margins ($\tau$) parameter.** The value of the uneven margin parameter is associated with the distribution of the positive and negative samples in the training data set. A value less than 1 is usually more appropriate for imbalanced training data set. Five values of 0~1 were tested, the results can also be seen in Table 4.

<table>
<thead>
<tr>
<th>Kernel</th>
<th>$\tau$</th>
<th>Precision</th>
<th>Recall</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear</td>
<td>1.0</td>
<td>0.9611278</td>
<td>0.94789934</td>
<td>0.9501495</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.9610505</td>
<td>0.94837993</td>
<td>0.9545919</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.9255299</td>
<td>0.9592452</td>
<td>0.94205445</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.8961971</td>
<td>0.96867436</td>
<td>0.93102384</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.7891829</td>
<td>0.98082125</td>
<td>0.87461305</td>
</tr>
<tr>
<td>Quadratic</td>
<td>1.0</td>
<td>0.9642115</td>
<td>0.92360836</td>
<td>0.9434393</td>
</tr>
<tr>
<td></td>
<td>0.8</td>
<td>0.9610784</td>
<td>0.9330375</td>
<td>0.94683105</td>
</tr>
<tr>
<td></td>
<td>0.6</td>
<td>0.94611204</td>
<td>0.93527466</td>
<td>0.9406523</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0.8814011</td>
<td>0.95956564</td>
<td>0.9188172</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0.77931607</td>
<td>0.97634697</td>
<td>0.8667749</td>
</tr>
</tbody>
</table>

Note: Feature set = {String, Orth, Root, Lookup}; context window size = 5.

From Table 4 we can see that quadratic kernel function yields slightly better precision than the linear kernel function for $\tau \geq 0.6$, but remains consistently lower recall for all the tested $\tau$ values. As a result, $F_1$ measure of linear kernel function achieves better result than that of the quadratic kernel function, which means the NER task may not benefit from higher order kernel functions. In addition, if $\tau = 0.8$, it yields the highest $F_1$ measure (0.9545919) and outperforms for both kernel functions at $\tau = 1.0$, although the improvement is not significant. This indicates that our data set might be slightly imbalanced. Based on the above results, Linear kernel
function and $\tau = 0.8$ are adopted as the optimal SVM model parameter settings in the following experiments.

### 6.2.3 Context window size

From the experiment, we want to learn which window size provides the most relevant context information helpful for recognizing the target named entity. Window sizes smaller than the default value 5 were tested and the results are shown in Table 5.

#### Table 5 System performance of the system with different context window sizes

<table>
<thead>
<tr>
<th>Context window size</th>
<th>Precision</th>
<th>Recall</th>
<th>$F_1$</th>
<th>Time (training)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.7535694</td>
<td>0.47531548</td>
<td>0.58259976</td>
<td>3,861 ms</td>
</tr>
<tr>
<td>1</td>
<td>0.9380531</td>
<td>0.954955</td>
<td>0.9464286</td>
<td>7,091 ms</td>
</tr>
<tr>
<td>2</td>
<td>0.95518357</td>
<td>0.9546107</td>
<td>0.9548696</td>
<td>8,030 ms</td>
</tr>
<tr>
<td>3</td>
<td>0.9586483</td>
<td>0.95189934</td>
<td>0.9552619</td>
<td>11,230 ms</td>
</tr>
<tr>
<td>4</td>
<td>0.9600417</td>
<td>0.948911</td>
<td>0.9544439</td>
<td>14,168 ms</td>
</tr>
<tr>
<td>5</td>
<td>0.9610505</td>
<td>0.94837993</td>
<td>0.9545919</td>
<td>16,637 ms</td>
</tr>
</tbody>
</table>

Note: Feature set = \{String, Orth, Root, Lookup\}; SVM linear kernel: $\tau = 0.8$.

Table 5 shows that without context information (context window size = 0), the system performance is much worse than those incorporating the context information. However, the larger context window size does not necessarily yield better overall system performance. Although the precision improves as the window size increases, and the recall decreases. In fact, by comparing $F_1$ measure, context window size = 3 achieves the best result, although the difference is not significant compared to the other window sizes greater than 1. This result could be explained by the fact that larger window may include more irrelevant context, thus adding noise to the feature vector which deteriorates the system performance. Moreover, larger window size results in greater dimension of the feature vector, which requires longer training and processing time. Hence, the trade-off between the system performance and efficiency also need be considered when choosing the context window size for practical applications. We choose context window size = 3 in the following experiments.

### 6.3 Experiments on recognizing different named entities

Using the parameter settings determined from the above experiments, the system performance on recognizing different named entities was further evaluated. Table 6 shows the average performance on each of 4 types of name entities required by the NER task.

As shown in Table 6, the overall system performance is promising for all types of name entities, as the precisions are all above 0.926775 and the $F_1$ measures are
Person-specific named entity recognition using SVM with rich feature sets

Hui NIE

Table 6  System performance on recognizing different types of named entities

<table>
<thead>
<tr>
<th>Entity type</th>
<th>Precision</th>
<th>Recall</th>
<th>( F_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree</td>
<td>0.9874128</td>
<td>0.9425025</td>
<td>0.96395284</td>
</tr>
<tr>
<td>Person</td>
<td>1.0</td>
<td>0.82297677</td>
<td>0.898149</td>
</tr>
<tr>
<td>Title</td>
<td>0.93590164</td>
<td>0.85178965</td>
<td>0.89176226</td>
</tr>
<tr>
<td>Organization</td>
<td>0.926775</td>
<td>0.8502885</td>
<td>0.8868006</td>
</tr>
</tbody>
</table>

Note: Feature set = \{{String, Orth, Root, Lookup}\}; SVM linear kernel: \( \tau = 0.8 \); context window size = 3.

all above 0.8868006. This indicates that the SVM classifier derived from the previous experiments is effective for solving the NER problem.

Among 4 types of named entities, recognition of “degree” achieves superior precision and recall, and hence the highest \( F_1 \) measure (0.96395284). This actually attributes to the fact that “degree” is a relatively simple entity with less number of possible patterns. Thus, the available training data is likely to include most (if not all) of the possible patterns, and the SVM classifier model is trained based on more complete information, therefore, more effective for recognizing the corresponding entity. The superior performance of recognizing “degree” entity also indicates that the selected NLP features in the experiments can fully represent the patterns of the “degree” entity.

In addition, it is interesting to note that the precision of “person” entity recognition is 1.0, which has by far exceeded the expectation, since none of name relevant lexicons have been utilized in the experiments. By further analyzing the feature vector composition, I believe the context words of a named entity and the semantic feature will play the key roles. In fact, person-specific Web pages usually include many commonly used words which are directly associated with person, such as Dr., professor, lecturer and etc. When such words are included as context and their semantic types are used in the feature set, they can provide sufficient evidences to ensure certain entity candidates relevant to “person” entity. However, as there was no enough training samples for learning the pattern of “person” entity, the recognition rules are incomplete, so the recall rate (0.82297677) is not as good as the precision rate. If more typical samples can be available for training, we would achieve more improvement on the recall rate.

7  Conclusions

In this article, an NER technique – an SVM based multi-classification approach combined with GATE and the current state-of-the-art NLP tool, is introduced which could effectively solve the NER problem in our IE task for person-specific information extraction.
This work shows that feature selection is the key issue in data-driven NER tasks and should receive great attention. The NLP-based features are of great importance; and semantic features are proved to be particularly beneficial for field-related NE extraction tasks.

Regarding SVM, the experiment results show that linear-kernel based model slightly outperforms quadratic-kernel based model, indicating the higher order kernel function may not be desired for the specific classification problem in practical application. In addition, model with uneven margin parameter ($\tau = 0.8$) outperforms the standard model ($\tau = 1.0$), suggesting that uneven margin SVM model would be a better approach for addressing imbalanced data.

Finally, investigation on different context window sizes shows that including context information significantly improves the system performance. However, larger window size does not necessarily yield better performance; and when choosing the optimal window size, the system efficiency should be taken into consideration as well.

In summary, the specially designed preliminary experiments make it feasible to fully explore the characters of the data and obtain the optimal parameter settings for the IE system, which leads to a preferable rate in recall, precision and $F_1$ measures. From the viewpoint of application, exploring the distinctive features of the target data is critical to find appropriate approaches to address the practical problems in data-driven applications.

In the future, further experiments will be conducted to systematically compare other start-of-the-art NER learning techniques (e.g. CRFs and HMM) for the person-oriented NER task, and apply them on multiple person-specific entity types. Moreover, the NER system is going to be incorporated into a complete information-extraction framework that mines attributes about people and relations among them from a large body of person-specific Web pages.

References


Person-specific named entity recognition using SVM with rich feature sets

Hui NIE


Reference Citation Format

The format for citations in text and for bibliographic references follows GB/T 7714—2005. The citation should be ordered in number as it appears in the text of the submitted article.

- **For journal article**

- **For book**

- **For proceedings**

- **For electronic journal article**

- **For electronic book**

- **For thesis or dissertation**
Submission Guidelines

♦ Aims
Chinese journal of Library and Information Science (CJLIS), being sponsored by the Chinese Academy of Sciences (CAS) and published quarterly by the National Science Library of CAS, is a scholarly journal in the field of library and information science (LIS). Its aim is to provide an international communication link between researchers, educators, administrators, and information professionals.

With the publication of the research results both from China and from other foreign countries, the Journal CJLIS strikes a balance between theory and practice. With its goal to provide an open forum for Chinese and international scholars in this field to exchange their research results, CJLIS also offers new possibilities in the advancement of Chinese library operations. The CJLIS tries to establish a platform for LIS students, researchers and library staff all over the world to engage in intellectual dialog and also to improve library services so as to promote even more quickened and substantial development of LIS in China.

♦ Scope
Striving toward academic excellence, innovation, and practicality, the CJLIS mainly includes research papers both on the theoretical as well as on the practical fronts in all aspects of the field. More specifically, it includes but not limited to informatics, library management, information technology application, knowledge organization system, knowledge management, archives, permanent preservation of library resources, LIS education, and so on.

♦ Refereeing Process
Articles and papers covering the topics or themes mentioned above, will be refereed through a double-blind peer review process.

♦ Editorial Advisory Board
The Editorial Board is composed of the nationally and internationally well-known scholars and researchers in the LIS field, the high quality of this Journal is thus reasonably assured.

♦ Manuscripts Categories
As the first English-language academic journal on LIS published in Mainland China, the CJLIS will take a proactive attitude to trace and report the prevailing hot issues in the field around the globe as well as the more serious scholarly communications. As such, the submitted manuscripts are classified into constant categories and unfixed categories. In the former category, research papers, library practice and progress reports are the essential components. In the latter, book reviews, biographical sketches, anecdotes, reminiscence of prominent librarians and brief communications will appear occasionally.

Research papers represent original research work or a comprehensive and in-depth analysis of a topic. More than 3,000 words are considered as a proper length for such manuscripts, with an abstract between 100 to 200 words.

Library practice covers the latest development and application in any segment of library field work and information service. The length of the manuscript is preferred to be more than 3,000 words, with an abstract between 100 to 200 words.

Progress reports reflect the projects result or research progress on the key topics of the library and information science. Submissions of articles to this section are expected to be comprehensive and analytical, which may deepen the understanding of the discussed issue and stimulate further researches on the topics, or give a new perspective on future technological applications. The manuscript length should be within 10,000 words, with an abstract between 100 to 200 words.

♦ Manuscripts Requirements
All papers can be submitted either in English or in Chinese (or both) with a double-line space. For the assurance that all the materials of the to-be-submitted are included, please check the following:

Title. Please give a brief biographical introduction to all contributing authors and their research background on a separated paper. For a better organization of the paper, please use the headings and subheadings.

Authors and affiliations. Please do not forget to write down the mailing address of each and every article contributors.

References. Be sure all the references used should be cited properly in both in-text and in bibliography. Particular attention should be paid to the proceedings. Do not forget adding the name(s) of editors of the compilation, as well as the name of the publishers. For the detailed information please request a copy of Reference Citation Format.

♦ Copyright
All submitted papers normally should not have been previously published nor be currently under consideration for publication elsewhere. For all the materials translated or obtained from other published resources, they should be properly acknowledged. All copyright problems should be cleared without any legal entanglements prior to the publication.

♦ Notes for Intending Submissions
A guide for authors and other relevant information, including submitting papers online, is available at the website of the Editorial Office of the CJLIS (http://www.chinalibraries.net). For any questions, you can e-mail the Office or directly to:

Prof. ZHANG Xiaolin
Editor-in-Chief of CJLIS
The CJLIS Editorial Office
National Science Library, Chinese Academy of Sciences
No.33 Beishuan Xilu, Zhongguancun, Haidian District,
Beijing 100190, P.R. China
Tel: 86-10-82624454 or 86-10-82626611 ext. 6628
Fax: 86-10-82624454
E-mail: chinalibraries@mail.las.ac.cn
Website: http://www.chinalibraries.net

♦ Subscription
For single-copy subscription in China: RMB ¥ 200/Issue. For subscription outside of China, US $ 199/Volume yearly (including air shipping)
Chinese Journal of Library and Information Science (Quarterly)

Contents

Research Papers
1 Defining an open access resource strategy for research libraries: Part III—The strategies and practices of National Science Library Xiaolin ZHANG, Xiwen LIU, Lin LI, Yan ZENG & Li-Ping KU

12 A view on big data and its relation to Informetrics Ronald ROUSSEAU

27 Person-specific named entity recognition using SVM with rich feature sets Hui NIE

47 A comparison of mapping strategies from DDC to CLC Fang LI & Yihua ZHANG

62 A study of information exchange through social networks in rural China Ya LIU

76 Is it time for wider acceptance of e-textbooks? An examination of student reactions to e-textbooks Ziming LIU

News

88 Reference Citation Format

Volume 5
Number 3
2012

ISSN 1674 – 3393
CN 11-5670/G2