THE NATIONAL UNIVERSITY
of SINGAPORE

School *of* Computing
Lower Kent Ridge Road, Singapore 119260

TRA6/03

Association Rules Mining for Name Entity Recognition

Indra Budi and Stephane Bressan

June 2003

Technical Report

Foreword

This technical report contains a research paper, development or tutorial article, which has been submitted for publication in a journal or for consideration by the commissioning organization. The report represents the ideas of its author, and should not be taken as the official views of the School or the University. Any discussion of the content of the report should be sent to the author, at the address shown on the cover.

JAFFAR, Joxan
Dean of School

Association Rules Mining for Name Entity Recognition

Indra Budi[1], Stéphane Bressan[2]
[1]Computer Science Faculty, University of Indonesia, indra@cs.ui.ac.id
[2]School of Computing, National University of Singapore, steph@nus.edu.sg

June 21, 2003

Abstract

We propose a new name entity class recognition method based on association rules. We evaluate and compare the performance of our method with the state of the art maximum entropy method. We show that our method consistently yields a higher precision at a competitive level of recall. This result makes our method particularly suitable for tasks whose requirements emphasize the quality rather than the quantity of results.

1 Introduction

The work presented in this report is part of a larger effort to develop information retrieval and linguistic systems, tools, and techniques for an Indonesian Digital Library (see [15, 14, 10]). The application that motivates this research requires that semantic structures in XML or RDF be extracted from and superimposed on a corpus of documents written in the Indonesian language. The natural first step of this project consists in identifying name entities from the texts in the corpus.

Name Entity Recognition (NER) is an information extraction task that is concerned with the recognition and classification of name entity from free text [9]. Name entities classes are, for instance, locations, person named, organization named, dates, and money amounts. To terms and express in the text correspond the entities they represent. For example, let see in the following sentence:

"British Foreign Office Minister O'Brien (right) and President Megawati pose for photographers at the Palace".

A name entity recognition process looking for named entities in that sentence would identify *O'Brien* and *Megawati* as named of person and the *Palace* as location named. This recognition can be based on a variety of features of the terms, the sentence, the text and its syntax and could leverage external sources of information such as thesauri and dictionaries, for instance. In the example, a system may have applied a simple rule guessing that the capitalized words directly following the terms 'President' or 'Minister' are names of persons.

In this report, we propose a method for name entity class recognition based on such rules in the form of association rules. The association rules defining the patterns of

syntax and term features potentially defining the classes are mined from a training set of documents.

The rest of this report is organized as follow. We present and discuss some background and related work on name entity class recognition in the next section. The method is presented in section 3 in which we detail the learning or mining phase and the testing and tagging phase. We evaluate and compare our method with the state of the art maximum entropy method [7] in section 4. Finally we conclude and identify the next steps of our research.

2 Background and Related Work

The first family of approaches to name entity recognition relies on the hand crafting of models and techniques for the recognition of entity classes. Generally the models consist of a set of patterns using grammatical (e.g. part of speech), syntactic (e.g. word precedence) and orthographic features (e.g. capitalization) in combination with dictionaries and thesauri. An example pattern in this type of system is the one we have suggested in our introductory example: "If a proper noun follows a person's title, then the proper noun is a person's name".

In this family of approaches Appelt et al. propose a name identification system based on carefully hand-crafted regular expression [2, 3], while Iwanska [11] uses extensive specialized resources such as gazetteers, and white and yellow pages. Morgan, for the same purpose, uses a highly sophisticated linguistic analysis [12]. These approaches are relying on manually coded rules and manually compiled corpora. They often yield prohibitive development and maintenance costs. Furthermore, for cost reduction and effectiveness reasons, they are often domain and language specific and do not necessarily adapt well to new domains and languages.

The alternative to hand-crafted approaches is the use of data mining, knowledge discovery and machine learning techniques. Both Sekine [13] and Bennett [4] propose name identification systems based on decision trees. The decision trees in both approaches use such features as part-of-speech, character, as well as dictionaries. Sekine's approach uses a single decision tree to compute the probability of a term to represent a given class while Bennet's approach combines multiple decision trees. Bikel in the popular Nymble system proposes a method based on the hidden markov model [5].

The maximum entropy approach to NER relies on the general maximum entropy technique for estimating probability distribution. Borthwick [7] described a word identification system built around a maximum entropy framework. The system uses a variety of knowledge sources, such as orthographic, lexical, section and dictionary features, to make tagging decisions. Chieu [8] uses maximum entropy and combines the local features we have mentioned before with global features in the text such as abbreviations in subsequent sentences.

3 Association Rule-based NER

In this section we described mining association rules and using of association rules for name entity recognition task.

3.1 Association Rules

Association rules and association rule mining [1] has received much attention in the last decade in the database and data mining community. The model and techniques have found many applications in a variety of domains. We mention for illustration results on Web caching [6] and on query expansion in information retrieval [16] by one of the author.

Association rule is a relationship of the form $X \Rightarrow Y$, where X and Y are sets of items from the dataset to be studied and $X \cap Y = \oslash$. Each association rule is assigned a support factor and a confidence factor.

$$support = \frac{X \cup Y}{N} \tag{1}$$

$$confidence = \frac{X \cup Y}{|X|} \tag{2}$$

Where N is the total number of records, $|A|$ denotes the number of records containing all the items in set A.

The support rule is the ratio of the number of items in X and Y over the total number of items; The confidence is the ratio of the number of items in X and Y over the number of items in X. The mining of association rules consists in extracting from the databases all such rules with support and confidence greater than or equal to a user-specified support and confidence.

3.2 Mining Association Rules for NER

In the name extraction task the datasets are documents which are sequences of terms with features and name classes. We use the set of features proposed by Bikel in [5] to which we added two features, *containsDigitsAndDollar* as monetary amount in dollar intuition and *containsDigitsAndColon* as time intuition. Table 1 shows the complete set of features we used. We consider the seven name classes considered in the standard MUC-7 benchmark (see [8]) for the test on MUC-7 (English) corpus: location named, person named, organization named, dates, times, monetary amount, and percentages. For Indonesian corpus we use three name classes which are location named, person named, and organization named.

The items we consider are occurrences of terms. However the sets X and Y can be described in terms of terms, sequences of terms, features and name classes. In practice Y is the name class we wish to predict. Among all the possible forms for X, after informal empirical tests, we settled to consider three types of rules.

Let us consider a sequence of terms $\langle t_1, t2 \rangle$, where f_2 is feature of t_2 and nc_2 is the name class of f_2. We consider the following three types of association rules:
1. $\langle t_2 \rangle \Rightarrow nc_2, (support, confidence)$
2. $\langle t_1, t_2 \rangle \Rightarrow nc_2, (support, confidence)$
3. $\langle t_1, f_2 \rangle \Rightarrow nc_2, (support, confidence)$

We called rules of type 1 dictionary rules, rules of type 2 bigram rules, and rules of type 3 feature rules.

Let us consider the example sentence *"Prof. Hasibuan conducted a lecture on information retrieval"*. In a training corpus in which name classes are given, the annotations of the corpus indicate that the term "Hasibuan" is name class of person.

Feature	Example text	Intuition
twoDigitNum	90	Two-digit year
fourDigitNum	1990	Four digit year
containsDigitAndAlpha	A8956-67	Product code
containsDigitAndDash	09-96	Date
containsDigitAndSlash	11/9/89	Date
containsDigitAndComma	23,000.00	Monetary amount
containsDigitAndPeriod	1.00	Monetary amount, percentage
containsDigitAndColon	12:30	Time
containsDigitAndDollar	$30	Monetary amount in dollar
otherNum	456789	Other number
allCaps	BBN	Organization
capPeriod	M.	Person name initial
initCap	Sally	Capitalized word
lowerCase	can	Uncapitalized word
other	,	Punctuation marks, all other words

Table 1: List of feature

We produce a dictionary rule of the form:
$\langle Hasibuan \rangle \Rightarrow person\_named(Hasibuan)$ with support and confidence depending on the number of occurrences of the term "Hasibuan" and the number of occurences of the term "Hasibuan" labelled in this entity class.

We produce a bigram rule of the form:
$\langle Prof., Hasibuan \rangle \Rightarrow person\_named(Hasibuan)$ with support and confidence depending on the number of occurrences of the expression "Prof. Hasibuan" and the term "Hasibuan" labelled in this name class.

We produce a feature rule of the form:
$\langle Prof., Capitalized\_word(X) \rangle \Rightarrow person\_named(X)$ with support and confidence depending on the occurrences of the expression "Prof. X" with X labelled in this feature and name classes

3.3 Using Association Rules for NER

The mined rules are considered for the name entity recognition task if their support and confidence is above user defined thresholds. We use the mined rules by type - dictionary, bigram, or feature- independently or combined. For every pair of terms in the text, the name entity recognition association rule-based algorithm, presented in figure 1, determines the rule with minimum support and highest confidence to be used. Ties, which are rare, are broken with a random choice. Not so rare is the case for which no rule is available. In that situation, the special class not-a-name is assigned to the term.

```
For every pair of terms <t1, t2>
   find the set R of rules X → nc
   such that <t1, t2> matches X
   and support and confidence are above threshold
   If R is not empty then
      Choose in R the rule X→ nc with highest confidence
      Assign nc as the name class of t2
   Else
      Assign not-a-name as name class of t2
Endfor
```

Figure 1: Recognition algorithm

We have experimented with various combinations of the rules, we report the results of five combinations of the three types of rules mined: the dictionary rules alone, the bigram rules alone, the combined bigram and dictionary rules, the feature rules alone and the combined feature and dictionary rules.

4 Experimental

4.1 Environments

In order to evaluate the effectiveness of our method we use the standard name entity recognition corpus, the MUC-7 corpus (see [8]). The corpus contains news articles in the English language in which terms representing name entities of seven classes (see section 3) have been labelled. We use a training set of 200 articles to learn the association rules and a testing set of 100 articles. The algorithm is implemented in C++.

We also ran experiments on a corpus of news articles in the Indonesian language. The corpus consists of 55 manually labelled news articles collected for eight days, between April 2nd 2001 and April 10th 2001, from the online version of the Indonesian newspaper Kompas (http://www.kompas.com) (see [15]). We took the articles vary in size from 201 to 1532 terms. As mentioned above, we use three name classes, person named, location named and organization named for experiment with Kompas corpus.

The effectiveness of the method is measured in terms of recall and precision. Recall is defined as the number of correct responses divided by number of answers. Precision is defined as the number of correct responses divided by the number of responses. A response is a term labelled by the algorithm with a name class. An answer is the name class of the term as labelled in the corpus. A response is correct if it corresponds to an answer.

In this series of experiment we compare our method with the maximum entropy

	Rule Association		Maximum Entropy	
	Recall	Precision	Recall	Precision
Dict	57.57	86.62	59.24	41.15
Bigram	34.37	93.21	57.40	65.03
Feature	44.84	67.75	49.56	58.99
Bigram+Dict	60.44	89.59	53.72	69.48
Feature+Dict	66.34	83.43	43.70	60.89
Bigram+Feature	53.73	77.61	59.61	76.10

Table 2: Result of experiment with MUC-7

	Rule Association		Maximum Entropy	
	Recall	Precision	Recall	Precision
Dict	48.57	77.73	48.29	77.80
Bigram	28.45	86.52	52.53	72.14
Feature	51.58	77.65	58.39	70.01
Bigram+Dict	48.42	82.61	78.09	70.20
Feature+Dict	62.80	81.35	63.10	61.13
Bigram+Feature	48.29	77.80	52.93	65.76

Table 3: Result of experiment with Kompas

method. For each type of association rules or combination of types of association rules that we use, we use the corresponding terms and features in the history of the maximum entropy method: one term for the dictionary method, two consecutive terms for the bigram method and a term and the feature of its successor for the feature method.

The maximum entropy method is implemented using the Java-based opennlp maximum entropy package (http://maxent.sourforge.net) on the same machine.

4.2 Results and Analysis

Experiments with MUC-7 corpus showed that the highest precision on association rules achieved by using bigram rules which is 93.21 % and the highest recall (66.34 %) achieved by using combination of feature and dictionary rule. On maximum entropy, the highest precision (76.10 %) and the highest recall (59.61 %) achieved by using combination bigram and feature. The details result with MUC-7 can be seen in Table 2.

For Indonesian corpus, highest precision (86.52) and recall (62.80) similar on association rules similar with MUC-7 corpus, achieved by bigram rules and combination of feature and dictionary rule respectively. On maximum entropy method, dictionary yielded the highest precision (77.80) and combined of bigram and dictionary yielded the highest recall (78.09). More details result with Kompas corpus can be seen in Table 3.

Table 2 and 3 showed that using bigram rule on association rules alone yields the highest precision with lower recall. Combination of feature and dictionary rule yielded the highest recall with competitive precision. On maximum entropy there is no consis-

```
<meeting>
  <date>05/06/2003</date format=europe>
  <location>
    <name>State Palace</name>
    <city>Jakarta</city>
    <country>Indonesia</country>
  </location>
  <participants>
    <person>
      <name>Megawati Soekarnoputri</name>
      <quality>President </quality>
      <country>Indonesia</country>
    </person>
    <person>
      <name>Mike O'Brien</name>
      <quality>Foreign Office
            Minister</quality>
      <country>Britain</country>
     </person>
  </participants>
</meeting>
```

Figure 2: Extracted XML data

tently results on using of history in MUC-7 and Kompas corpus. In MUC-7 corpus seem using combination bigram and feature as history yield the highest recall and precision but in Kompas corpus dictionary as history and its combined with bigram yield the highest precision and recall.

The Association rule based methods consistently yields a higher precision than the corresponding maximum entropy based method on experiments with English and Indonesian corpus.

The dictionary rules can be used in combination with the bigram or the feature rules to significantly increased the recall (namely many terms occur without any particular contextual features). The various combinations of features also seem to impact consistently in both association rule and maximum entropy based methods.

For the Indonesian corpus, with only three name entity classes, the association rule based methods have difficulties to reach a competitive level of recall.

5 Conclusions and Future Work

We have presented a new name entity class recognition method based on association rules. We compared our method with the maximum entropy method. Our experiments showed that association rules consistently yield a higher precision than the maximum entropy method. In the English corpus, under the appropriate combination of types of rules it is possible to improve the recall so that the association rule method is strictly more effective than the maximum entropy.

The next step in our project is to devise a method to reconstruct structured elements

from the elementary name entities identified. Our target language is XML. To illustrate our idea, let us consider the motivating example from which we wish to extract an XML document describing the meeting taking place: "British Foreign Office Minister O'Brien (right) and President Megawati pose for photographers at the State Palace." Figure 2 contains the manually constructed XML we hope to obtain. In italic are highlighted the components that require global, ancillary, or external knowledge. Indeed, although, we expect similar methods (association rules, maximum entropy) can be used to learn the model of combination of elementary entities into complex elements, we also expect that global, ancillary, and external knowledge will be necessary such as lists of names of personalities (Mike O'Brien, Megawati Soekarnoputri), gazetteers (Jakarta is in Indonesia), document temporal and geographical context (Jakarta, 05/06/2003), etc.

References

[1] Agrawal, R., Tomasz Imielinski, and Arun Swami, *Mining association rules between sets of items in large databases*, Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data (Washington, D.C.), ACM, 1993.

[2] Appelt, D. and et. al., *Fastus: A finite state processor for information extraction from real-world text*, Proceedings of IJCAI 1993, 1993.

[3] Appelt, D. and et. al., *Sri international fastus system muc-6 test results and analysis*, Proceedings of the Sixth Message Understanding Conference (MUC-6) (Columbia), NIST, Morgan-Kaufmann Publisher, 1995.

[4] Bennett, S.W., Chinatsu Aone, and Craig Lovell, *Learning to tag multilingual texts through observation*, 1997, pp. 109–116.

[5] Bikel, D., S. Miller, R. Schwartz, and R. Weischedel, *Nymble: A high performance learning name-finder*, Proceeding of the fifth Conference on Applied Natural Language Processing (Washington, D.C.), ACL, 1997, pp. 194–201.

[6] Bin, L., S. Bressan and B.C. Ooi, *Making web servers pushier*, Proceedings of the Workshop on Web Usage Analysis and User Profiling Springer Verlag, August 1999., 1999.

[7] Borthwick, A., J. Sterling, E. Agichtein, and R. Grishman, *Exploiting diverse knowledge sources via maximum entropy in named entity recognition*, 1998.

[8] Chieu, H.L., and Hwee Tou Ng, *Named entity recognition: A maximum entropy approach using global information*, 2002.

[9] Grishman, Ralph, *Information extraction: Techniques and challenges*, 1997, Lecture Notes in Computer Science, Vol. 1299, Springer-Verlag.

[10] Indradjaja, L. and S. Bressan, *Automatic learning of stemming rules for the indonesian language*, Proceeding of the The 17th Pacific Asia Conference on Language Information and Computation (PACLIC), 2003.

[11] Iwanska, L., M. Croll, T. Yoon and M. Adams, *Wayne state university: Description of the uno natural language processing system as used for muc-6*, Proceedings of the Sixth Message Understanding Conference (MUC-6) (Columbia), NIST, Morgan-Kaufmann Publishers, 1995.

[12] Morgan, R. and et. al., *University of durham: Description of the lolita system as used for muc-6*, Proceedings of the Sixth Message Understanding Conference (MUC-6) (Columbia), NIST, Morgan-Kaufmann Publishers, 1995.

[13] Sekine, S., R. Grishman, and H. Shinnou, *A decision tree method for finding and classifying names in japanese texts*, Proceedings of the Sixth Workshop on Very Large Corpora, Montreal, Canada, 1998.

[14] Vinsensius, V. and Bressan, S., *Continuous-learning weighted-trigram approach for indonesian language distinction: A preliminary study*, Proceedings of 19th International Conference on Computer Processing of Oriental Languages, 2001.

[15] Vinsensius, V. and Bressan, S., *Temu-kembali informasi untuk dokumen-dokumen dalam bahasa indonesia*, Electronic proceedings of Indonesia DLN Seminar, 2001.

[16] Wei, J., Qin, Z., Bressan, S. and B.C. Ooi, *Mining term association rules for automatic global query expansion: A case study with topic 202 from trec4*, Proceedings of the Americas Conference on Information Systems 2000 Data Mining and Information Retrieval in Business, 2000.