Vietnamese Word Segmentation with CRFs and SVMs: An Investigation

Cam-Tu Nguyen[1], Trung-Kien Nguyen[1], Xuan-Hieu Phan[2]
Le-Minh Nguyen[2], and Quang-Thuy Ha[1]

[1] College of Technology, Vietnam National University, Hanoi
[2] School of Information Science, Japan Advanced Institute of Science and Technology
ncamtu@vnu.edu.vn , ntkien@vnu.edu.vn , hieuxuan@jaist.ac.jp
nguyenml@jaist.ac.jp , thuyhq@vnu.edu.vn

Abstract. Word segmentation for Vietnamese, like for most Asian languages, is an important task which has a significant impact on higher language processing levels. However, it has received little attention of the community due to the lack of a common annotated corpus for evaluation and comparison. Also, most previous studies focused on unsupervised-statistical approaches or combined too many techniques. Consequently, their accuracies are not as high as expected. This paper reports a careful investigation of using conditional random fields (CRFs) and support vector machines (SVMs) - two of the most successful statistical learning methods in NLP and pattern recognition - for solving the task. We first build a moderate annotated corpus using different sources of materials. For a careful evaluation, different CRF and SVM models using different feature settings were trained and their results are compared and contrasted with each other. In addition, we discuss several important points about the accuracy, computational cost, corpus size and other aspects that might influence the overall quality of Vietnamese word segmentation.

Keywords: Word segmentation, segmenting and labeling sequence data, conditional random fields, support vector machines, maximum matching.

1 Introduction

Word segmentation is one of the fundamental preprocessing steps in NLP for building higher applications. It is even more important and challenging in Asian languages, such as Chinese, Japanese, and Korea, because there is no white space between two consecutive words. Vietnamese language faces a similar problem due to the fact that a word may contain more than one separated syllables, and therefore the white space is not always the word separator.

In recent years, word boundary detection for Vietnamese has received more attention from the community and there have been several statistical and machine learning methods applied to the task. However, most of the current methods either suffer from unsatisfactory results [1, 2] (with accuracy of 91% or lower) or must combine many techniques in a multi-level processing to obtain good results [3]. In addition, their works were done without comparison to any baseline or study of the quality of the corpus (e.g., the out-of-vocabulary rate, the number of date/time and numbers). Also, there is still no common standard annotated corpus for evaluation and comparison.

In this paper, we present a thorough investigation of using two powerful statistical learning methods, CRFs and SVMs, to perform the task. To do so, we first build an annotated corpus of about 8000 sentences with word boundary marked. Although the corpus is not large enough to cover a broad range of Vietnamese vocabularies, it contains documents from different domains to reduce the imbalance in word distribution. Then, CRF and SVM models are trained on the corpus using various feature configurations and their performances are compared and contrasted with each other to determine the impact of feature selection as well as the generalization power of CRFs and SVMs on the segmentation accuracy. CRF and SVM models are also compared with a baseline (maximum matching from a Vietnamese dictionary) to see the extent to which machine learning techniques can help to improve the

accuracy in comparison with the simple heuristics-based matching approach. All in all, our main motivations behind this work can be summed up as follows:

- CRFs and SVMs have recently been seen as two of the most successful statistical learning models in NLP and pattern classification. While SVM together with appropriate kernels is well-known thanks to its optimal discriminating hyper-plane in very high-dimensional feature spaces, CRF is particularly designed for problems of labeling and segmenting sequence data because of its global normalization and trade-off among state variables of sequence data. In addition, both SVMs and CRFs follow the discriminative learning approach and have a big flexibility to encode a variety of features from the input data to enhance their the prediction capability. Recently, both SVMs and CRFs have been applied to a wide range of labeling and segmenting tasks in NLP like POS tagging, chunking, named entity recognition, and information extraction, and have achieved state-of-the-art results. Therefore, we do hope that CRFs and SVMs can achieve similar successes in segmenting Vietnamese.
- Feature selection is a core step in both CRFs-based and SVMs-based approaches. The important issue is that which types of feature should be used and how they influence the accuracy of the system. Note that chosen features should be useful to the word segmentation task and appropriate to Vietnamese. Because Vietnamese word segmentation is still an open task, a careful estimation of feature types especially ones bearing characteristics of Vietnamese will set up a foundation for future research on Vietnamese word segmentation.

The remaining part of the paper is organized as follows. Section 2 briefly discusses word formation in Vietnamese. Section 3 describes quickly the two learning models CRFs and SVMs for labeling and segmenting sequence data. Section 4 mainly presents the framework of using CRFs and SVMs for Vietnamese word segmentation. Corpus building, feature selection, result comparison and analysis will be presented in this section. Finally, some conclusions will be given in Section 5.

2 Vietnamese Word Formation

2.1 Vietnamese Syllable

Vietnamese syllables are elementary units that have one way of pronunciation. In documents, they are usually delimited by white-space. Being the elementary units, Vietnamese syllables are not undivided elements but a structure [6]. Table 1 depicts the general structure of Vietnamese syllable

Table 1. Structure of Vietnamese Syllable

TONE MARK			
First Consonant	Rhyme		
	Secondary Consonant	Main Vowel	Last Consonant

Generally speaking, each Vietnamese syllable has all five parts: first consonant, secondary vowel, main vowel, last consonant and a tone mark. For instance, the syllable “tuần” (week) has a tone mark (grave accent), a first consonant (t), a secondary vowel (u), a main vowel (â) and a last consonant (n). However, except for main vowel that is required for all syllables, the other parts may be not present in some cases. For example, the syllable “anh” (brother) has no tone mark, no secondary vowel and no first consonant. In other case, the syllable “hoa” (flower) has a secondary vowel (o) but no last consonant.

2.2 Vietnamese Word

A word in Vietnamese may consist of one or more syllables which are combined in different ways. Based on the way of constructing words from syllables, we can classify them into three categories: single words, complex words and reduplicative words.

- Simple Words: a single word has only one syllable that implies a specific meaning. For example: tôi(I), bạn(you), nhà(house), ...
- Complex Words are words that consist of more than one syllable. The syllables in a complex word are combined based on semantic relationships, that is either coordinated (bơi lội– swim) or "principle and accessory" (đường sắt – railway).
- A word is considered as a reduplicative one if its syllables have phonic components reduplicated. Reduplicative words are usually used for scene or sound descriptions, thus they usually used in literary texts.

2.3 New Words in Vietnamese

New words are those that are out-of-vocabulary or not present in a particular dictionary or training corpus. In our system, we consider 3 types of new words: abbreviations, named entities, and foreign words.

- Abbreviations are usually all-capitalized, for example: CAND (Công An Nhân Dân - people's police).
- Named entities are proper nouns indicating person, location, organization, date-time and so on. In Vietnamese, personal or organizational names often have first letter capitalized, for instance: "Hồ Chí Minh" or "Hải Hà" in phrase "Công ty Hải Hà" (Hải Hà firm). The other named entities such as number, percentage, date-time, etc. can also be captured by some specific regular expressions.
- Foreign words used in Vietnamese are usually Latinized. Each foreign word has only one syllable that does not often conform to the structure of Vietnamese syllable described in section 2.1.

By observing types of new words in documents, we later introduce some kinds of useful features in order to detect them in order to increase the accuracy of our system.

3 Sequential Labeling with CRFs and SVMs

3.1 Conditional Random Fields

In this paper, CRFs are referred to as an undirected linear-chain of model states, i.e., a conditionally-trained finite state machine (FSMs) that obey the first-order Markov property.

Let $\mathbf{o} = (\mathbf{o_1}, \mathbf{o_2}, \ldots, \mathbf{o_T})$ be some observed data sequence. Let S be a set of FSM states, each of which is associated with a label $l \in L$. Let $\mathbf{s} = (\mathbf{s_1}, \mathbf{s_2}, \ldots, \mathbf{s_T})$ be some state sequence, CRFs [4] define the conditional probability of a state sequence given an observation sequence as

$$p_\theta(\mathbf{s} \mid \mathbf{o}) = \frac{1}{Z(\mathbf{o})} \exp\left[\sum_{t=1}^{T}\sum_{k} \lambda_k f_k(s_{t-1}, s_t, \mathbf{o}, t)\right]. \quad (1)$$

Where $Z(\mathbf{o}) = \sum_{s'} \exp\left(\sum_{t=1}^{T}\sum_{k} \lambda_k f_k(s'_{t-1}, s'_t, \mathbf{o}, t)\right)$ is normalization summing over all label sequences. f_k denotes a feature function in the language of maximum entropy modeling and λ_k is a learned weight associated with feature f_k. Each f_k is either a *per-state* or a *transition* feature.

$$f_k^{(per-state)}(s_t, \mathbf{o}, t) = \delta(s_t, l) x_k(\mathbf{o}, t). \quad (2)$$

$$f_k^{(transition)}(s_{t-1}, s_t, t) = \delta(s_{t-1}, l)\delta(s_t, l). \quad (3)$$

Where δ denotes the Kronecker-δ. A per-state feature (2) combines the label l of current state s_t and a context predicate, i.e., the binary function $x_k(\mathbf{o},t)$ that captures a particular property of the observation sequence **o** at time position t. For example, the current label is B_PER and the current word is "Nguyễn". A transition feature (3) presents sequential dependencies by combining the label $l'$of the previous state s_{t-1} and the label l of the current state s_t, such as the previous label l'=B_PER and the current label is l=I_PER.

Training CRFs is commonly performed by maximizing the likelihood function with respect to the training data using advanced convex optimization techniques like L-BFGS. And inference in CRFs, i.e., searching the most likely output label sequence of an input observation sequence, can be done using Viterbi algorithm.

3.2 Support Vector Machine

Suppose that we have a set of training samples $\mathbf{D} = (\mathbf{x_1}, \mathbf{y_1}), \ldots, (\mathbf{x_l}, \mathbf{y_l})$ ($\mathbf{x_i} \in R_n$, $\mathbf{y_i} \in \{+1, -1\}$) where $\mathbf{x_i}$ is a feature vector of the i-th sample represented by an n dimensional vector, $\mathbf{y_i}$ is the class (positive(+1) or negative(-1)) label of the i-th sample. l is the number of training samples. The main ideas behind SVMs is to separate positive and negative samples by a hyperplane expressed as $(\mathbf{w}.\mathbf{x}) + b = 0$. SVMs [11] find the separating hyperplane which maximizes its margin. In other words, this problem becomes equivalent to solving the following optimization problem:

Maximize: $\boldsymbol{M}=2/\|\mathbf{w}\|$

Subject to: $\mathbf{y_i}[(\mathbf{w}.\mathbf{x_i}) + b] \geq 1$. Not only the linear classification, SVMs also can carry out the non-linear ones by introducing kernel functions that embed the data into a feature space where the nonlinear pattern now appears linear. Though, we omit the details here, the key aspect of kernel functions is that they preserve the pairwise inerproducts while relaxing the constraints of coordinates of the embedded points.

Basically, SVMs are binary classifier, thus we must extend SVMs to multi-class classifier in order to classify three or more classes. The *pairwise* classifier is one of the most popular methods to extend the binary classification task to that of $\boldsymbol{K}$ classes. Though, we leave the details to [12], the idea of *pairwise* classification is to build $\boldsymbol{K.(K-1)}/2$ classifiers considering all pairs of classes, and final decision is given by their weighted voting.

4 Vietnamese Word Segmentation with CRFs and SVMs

4.1 Corpus Building

Building a robust and accurate word segmentation system for Vietnamese using machine learning approaches is more complex in comparison with building such a system in other languages due to the fact that there is no standard corpus publicly available. For this work, we have to build a corpus of 305 newspaper articles from many websites and in various domains. Although the corpus is not large enough to cover a broad range of Vietnamese words, it contains documents from multiple domains with the hope that this will reduce the imbalance in word distribution. This corpus is now available online[1].

In addition to the corpus, we also collect other resources that are used as lexicons or domain-knowledge for our system: a Vietnamese dictionary, a list of 2000 personal names[2], and a list of 707 names of locations[3] in Vietnam. These can be seen as external dictionaries and will be used for looking up in our CRF and SVM models.

[1] http://www.jaist.ac.jp/~hieuxuan/vnwordseg/data
[2] From http://www.vietnamgiapha.com/
[3] From http://vi.wikipedia.org/

Table 2. Statistics of our annotated corpus for Vietnamese word segmentation

No	Domain	Number of Documents
1	Economics	90
2	Information Technology	59
3	Education	38
4	Vehicle	35
5	Sports	28
6	Law	31
7	Culture-Society	24
Total	305 newspaper articles (about 7800 sentences)	

4.2 Problem Representation

We cast the segmentation problem as a sequential tagging task: Vietnamese syllable that begins a word is marked with B_W (Begin of a Word), the syllable that is inside a word is marked with I_W (Inside of a Word), and the other things like comma, dot marks are tagged with O (Outside of a word). The problem of detecting word boundaries in a sentence is modeled as the problem of labeling syllables in that sentence with three above labels.

Performance of both the CRFs-based and SVMs-based segmentation systems depends on how well we do feature selection. In the following section, we will discuss more about the strategies of feature selection for both CRFs and SVMs.

Feature Selection

As mentioned earlier, there are two kinds of feature functions used in linear-CRFs: edge features which obey to the first-Markov property, and per-state features which are generated by combining information (context predicate) available surrounding current position in the observation sequence with the current label. Based on the same idea, we also integrate 2 kinds of features into SVM model. They are static features and dynamic features. While SVM model decide dynamic features in tagging process by considering the two previous labels, static features are very similar to per-state features in CRF model in the sense that we also take into account context predicates at the current observation.

After studying specific characteristics of Vietnamese, we propose five types of context predicate templates from which various features will be generated correspondingly.

Table 3. Context predicate temlates for CRFs and SVMs. Here, a valid Vietnamese syllable must conform to the structure described in section 2; Is_Marks will check current observation is a full stop, comma or some kinds of marks in the sentence or not; Is_Regular_Expression tries to capture expressions describing date/time (long date/ short date), numbers or tokens that is a mix of characters and numbers, etc.

Syllable Conj. (*SC*)	Syllable_Conjunction (-2,2)
Dictionary (*Dict*)	In_LacViet_Dictionary (-2,2)
External Resources (*ERS*.)	In_Personal_Name_List(0,0), In_Family_Name_List(0,0), In_Middle_Name_List(-2, 2), In_Location_List(-2,2)
Miscellaneous (*Misc*)	Is_Regular_Expression(0,0), Is_Initial_Capitalization(0,0), Is_All_Capitalization(0,0), Is_First_Observation(0,0), Is_Marks(0,0)
Vietnamese Syllable Detection (*VSD*)	Is_Valid_Vietnamese_Syllable(0,0)

Table 3 summarizes the context predicate templates used in our models. Note that, two numbers inside the brackets next to each context predicate template indicate the window surrounding current position in which we explore context information. For examples, In_LacViet_Dictionary (-2, 2) means that we make a particular conjunction of adjacent syllables in the sliding window from the second previous to the second next syllables and check if that conjunction forms a word in the dictionary. Also,

because the number of more-than-4-syllables words is quite low, we take into account only conjunctions of up-to-three adjacent syllables. Similarly, by Syllable_Conjunction (-2, 2), we considers all the 1-gram, 2-gram and 3-gram of syllables in the window of size 5.

4.4 Experimental Setup

Our experiments with CRFs and SVMs were conducted using two tools: FlexCRFs - a C/C++ implementation of CRFs, and Yamcha - a C/C++ implementation of SVMs for labeling and segmenting sequence data. For CRF models, we use first-order Markov dependency, and for SVM models we use the second degree of polynomial kernel.

Table 4. Sumary of our experiment design. Here, SC stands for Syllable Conjunction, Misc: Miscellaneous, Dict: Dictionary, ERS: External Resources, VSD: Vietnamese Syllable Detection

ID	Feature settings for experiments
1	Maximum Matching (looking up a Vietnamese dictionary)
2	CRFs with SC
3	CRFs with SC + VSD
4	CRFs with SC+VSD+ Dict
5	CRFs with SC+ VSD+ Dict + ERS
6	CRFs with SC+ VSD + Dict + ERS + Misc
7	SVMs with SC
8	SVMs with SC+VSD
9	SVMs with SC+ VSD+ Dict

For evaluating in each experiment, we used 5-fold cross-validation test. In other words, we randomly divide the corpus into five partitions; in each fold we take one partition as the testing set and all the others for training. Additionally, for the purpose of evaluating the performance of new word detection process, the OOV (out-of-vocabulary) rate is measured for 5 folds in the corpus and the average result among 5 folds is about 11.6%. This OOV rate is quite high in comparison with some popular corpora for Chinese word segmentation such as the corpora of Beijing University (6.9%), Hong Kong City U (7.1%) and Academia Sinica (2.2%) [8].

4.5 Experimental Results and Discussion

One of the most noticeable advantages of CRFs is the flexibility of integrating various kinds of features from the training data, thus we design 5 separate experiments for CRFs - the later is richer of features than the earlier – to estimate the effect of feature types on the system performance. For SVMs-based segmentation, we conduct an experiment with the strategy of feature selection described in the section 4.2. Though this work is an investigation of using CRFs and SVMs for Vietnamese Word Segmentation, we still need a baseline for all experiments and we have used Maximum Matching for this purpose. The summary of the design for our experiments is showed in Table 4

The average precision, recall, F1-measure of 5-fold CV tests from 9 experiments are summarized in Table 5. Figure 1 depicts the performance comparison among CRF models using different feature settings (Maximum Matching as a baseline). Also, the performance comparison between SVM and the CRF models (with two similar feature settings) is shown in Figure 2.

From Table 5 and Figure 1, several observations could be made. First, even with the simplest feature setting, CRFs can outperform Maximum Matching significantly. This is because the large number of new words appears in the data and partly shows that our corpus has high out-of-vocabulary rate. Second, the richer feature setting is, the better results CRFs model can achieve. The CRF model with richest feature setting yields the highest performance (average F_1 of 94.05%) in comparison with the other settings (average F_1 of 93.98%, 93.90%, 90.41% and 90.14%). Third, the impact of different feature types on the performance of CRFs model is different. As we can see in the figure, the Dict and SC feature types have the largest effect on the performance of the system. While the Misc feature type only

has a little effect. This is due to several reasons: (1) Vietnamese word boundary depends more on identity of its syllable than their features like capitalization or that it is the first syllable in a sentence or not. (2) Though regular expressions can detect number, date/time quite well, the number of date/time and number expressions is much smaller than the number of words in the corpus.

Table 5. The average precision, recall, and F1-measure of 5-fold cross-validation tests of Maximum Matching, CRFs and SVMs

Method (feature settings)	**Pre (%)**	**Re (%)**	**F_1**
Maximum Matching	87.65	78.59	82.82
CRFs (SC)	89.98	90.30	90.14
CRFs (SC+VSD)	90.24	90.58	90.41
CRFs (SC+VSD+Dict)	93.67	94.12	93.90
CRFs (SC + VSD + ERS)	93.71	94.26	93.98
CRFs (SC + VSD + ERS + Misc)	**93.76**	**94.28**	**94.05**
SVMs (SC)	90.60	91.45	91.02
SVMs (SC + VSD)	90.21	90.87	90.54
SVMs (SC + VSD + Dict)	**94.00**	**94.45**	**94.23**

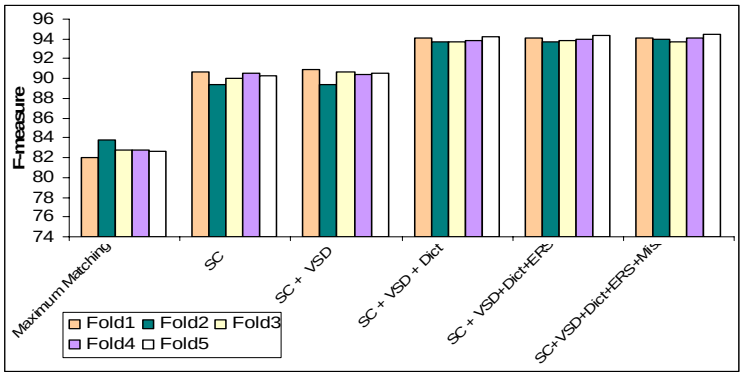

Fig. 1. Performance comparison among CRF models using different feature settings (Maximum Matching as the baseline)

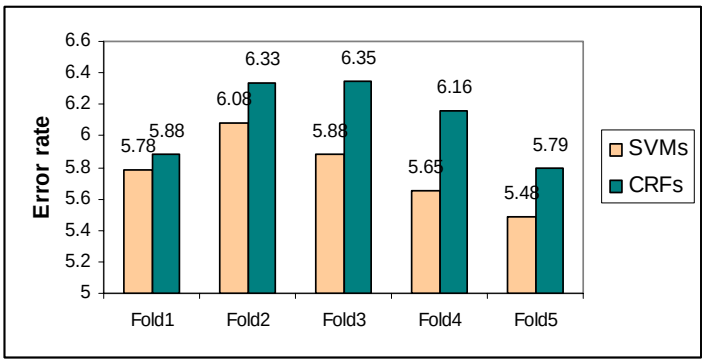

Fig. 2. Performance comparison between CRFs (SC + VSD + Dict) and SVMs on 5-fold CV test

Table 5 and Figure 2 show us some interesting information. Although VSD feature enhances the CRF model a little bit, it does not have positive effect on the SVM model. This shows that inappropriate features may cause noisy in the model and result in worse performance. Another observation is that with the similar feature settings, the SVM model is better than the similar CRF model. Moreover, SVM model with (SC+VSD+Dict) even performs better than the CRF model with the richest feature setting

(the average F_1 of 94.05%). To further study the two methods, we later compare computational time of SVMs and CRFs and see that with the similar feature setting, SVM model is slower than CRF model. For example: with the feature setting SC+VSD+Dict, the SVM model took 4 hours to complete the training process for 1 fold but the CRF model took only 2 hours for the same task. However, if comparing this SVM model to CRF model with richest feature setting (SC+VSD+Dict+ERS+Misc), we see they need nearly the same computational time while the SVM model is a little bit better than the CRF model in performance. This shows that SVM-based learning is a potential approach to the problem of Vietnamese segmentation. With good feature selection, it gives a little bit better performance than CRFs in this task.

5 Conclusions

We have presented a thorough investigation of using CRFs and SVMs for Vietnamese word boundary detection. The key contributions of our work is three-fold: (1) building an annotated corpus for evaluation; (2) training different CRF and SVM features according to different feature configurations; and then compare and contrast their results; and (3) draw some interesting conclusions that we observed from the experimental results.

Due to the computational burden of SVMs and the time limitation, the complete experiments for SVMs (of all feature settings) will be done in the near future. Also, in future work we need to evaluate how segmentation accuracy depends on the corpus size.

Acknowledgments. This work is partly supported by Project QC.06.07 ``Fundamental Vietnamese Shallow Processing with Modern Statistical Machine Learning Models'', Vietnam National University, Hanoi, and Project ``Information Extraction Models for finding Entities and Semantic Relations in Vietnamese Web Pages'' of the Ministry of Science and Technology, Vietnam.

References

1. Ha, L.A.: A method for word segmentation in Vietnamese. Corpus Linguistics, Lancaster, UK (2003)
2. Nguyen, T.V., Tran, H.K., Nguyen, T.T.T., Nguyen, H.: Word segmentation for Vietnamese text categorization: an online corpus approach. Research, Innovation and Vision for the Future, The 4th International Conference on Computer Sciences (2006)
3. Dinh, D., Kiem, H., Toan, N.V.: Vietnamese Word Segmentation. The 6th Natural Language Processing Pacific Rim Symposium (2001), 749--756s
4. Lafferty, J., Mccallum, A., Pereira, F.: Conditional Random Fields: probabilistic models for segmenting and labeling sequence data. The 18th International Conference on Machine Learning, Massachusetts, USA (2001), 282—290
5. Liu, D.C., Nocedal, J.: On the limited memory BFGS method for large scale optimization. Mathematical Programming 45 (1989) 503--528.
6. Mai, N.C., Vu, D.N., Hoang, T.P.: Foundations of linguistics and Vietnamese. Education Publisher (1997) 142-152
7. Nguyen, H., Nguyen, H., Vu, T., Tran, N., Hoang, K.: Internet and Genetics Algorithm-based Text Classification for Documents in Vietnamese. Research, Innovation and Vision for the Future, The 3th International Conference on Computer Sciences (2005)
8. Peng, F., Feng, F., McCallum, A: Chinese Segmentation and New Word Detection using Conditional Random Fields. The 20th International Conference on Computational Linguistics (2004).
9. Rabiner: A tutorial on hidden markov models and selected applications in speech recognition. In Proc. the IEEE, 77(2):257-286, 1989.
10. Sha, F., Pereia, F.: Shallow parsing with Conditional Random Fields. Human Language Technology (2003)
11. Vapnik, V.N.: Statistical Learning Theory, Wiley-Interscience
12. Kudo, T., Matsumoto, Y.: Chunking with Support Vector Machines, The Second Meeting of the North American Chapter of the Association for Computational Linguistics (2001)