Exploration of N-gram Features for the Domain Adaptation of Chinese Word Segmentation

Zhen Guo, Yujie Zhang, Chen Su, and Jinan Xu

School of Computer and Information Technology,
Beijing Jiaotong University, Beijing 100044, China
{08281153,yjzhang,08281138,jaxu}@bjtu.edu.cn

Abstract. A key problem in Chinese Word Segmentation is that the performance of a system will decrease when applied to a different domain. We propose an approach in which n-gram features from large raw corpus are explored to realize domain adaptation for Chinese Word Segmentation. The n-gram features include n-gram frequency feature and AV feature. We used the CRF model and a raw corpus consisting of 1 million patent description sentences to verify the proposed method. For test data, 300 patent description sentences are randomly selected and manually annotated. The results show that the improvement of Chinese Word Segmentation on the test data achieved at 2.53%.

Keywords: Chinese Word Segmentation, CRF, domain adaptation, n-gram feature.

1 Introduction

Chinese Word Segmentation (CWS) methods can be roughly classified into three types: dictionary-based methods, rule-based methods and statistical methods. Due to the large number of new words appearing constantly and the complicated phenomena in Chinese language, the expansion of dictionaries and rules encounter a bottleneck. The first two methods are therefore difficult to deal with the changes in language usages. Since the statistical method can easily learn new words from corpora, Chinese Word Segmentation systems based on such methods can achieve high performance. For this reason, the statistical method is strongly dependent on annotated corpora. Theoretically, the larger amount of the training data with higher quality annotation will bring about the better performance for Chinese Word Segmentation systems.

Texts to be processed may come from different domains of the real world, such as news domain, patent domain, and medical domain, etc. When switching from one domain to another, both vocabulary and its frequency distribution in texts usually vary because the ways of constructing words from characters are different. This fact brings great challenges to Chinese Word Segmentation. The accuracy of a developed system trained on one domain will decrease obviously when it is applied to the texts from other different domains.

M. Zhou et al. (Eds.): NLPCC 2012, CCIS 333, pp. 121–131, 2012.

A solution is to develop domain-specific system for each domain by using corresponding annotated data. In application, the target domain of a text to be processed is recognized first and then the corresponding system is applied. In this way, the best results can be expected. However, manual annotation is a time-consuming and hard work. At present, a large amount of annotated data with high quality is not available for each domain and therefore it is not practical to develop domain-specific system in this way.

In order to solve the problem of the domain adaptation for Chinese Word Segmentation, many methods have been proposed, such as data weighting algorithm and semi-supervised learning algorithm. (Meishan Zhang et al., 2012) propose an approach that can incorporate different external dictionaries into the statistical model to realize domain adaptation for Chinese Word Segmentation. However, these methods have limitations because the annotated corpora and domain-specific dictionaries are not resources readily available.

This paper proposes an approach in which n-gram features from large raw corpus are explored to realize domain adaptation for Chinese Word Segmentation. Compared with annotated corpus and domain-specific dictionary, a raw corpus is more easily to obtain. Our experimental results show that the proposed approach can effectively improve the ability of the Chinese Word Segmentation system for domain adaptation.

2 Chinese Word Segmentation Based on Conditional Random Fields

Conditional random fields (CRF) is a probabilistic framework for labeling and segmenting sequential data, which is first proposed by (John Lafferty et al.,2011) on the basis of the maximum entropy models and hidden Markov models. CRF are undirected graphical models in which the parameters are estimated by maximizing the joint probability over observation and label sequences given an observation sequence. Linear-chain CRF is most widely used in machine learning tasks.

2.1 Conditional Random Fields

A linear-chain CRF with parameters $\Lambda= (\lambda_1,\lambda_2,\ldots,\lambda_n)$ defines a conditional probability for a label sequence $Y= (y_1,y_2,\ldots,y_n)$ given an input sequence $X= (x_1,x_2,\ldots,x_n)$ to be:

$$P_{\Lambda}(Y \mid X) = \frac{1}{Z_X} \exp\{\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(y_{t-1}, y_t, X, t)\} \tag{1}$$

Where $f_k (y_{t-1}, y_t, X, t)$ is a feature function which is often binary-valued, λ_k is a learned weight associated with feature f and Z_X is the normalization factor over all state sequences.

$$Z_X = \sum_y \exp\{\sum_{t=1}^{T} \sum_k \lambda_k f_k(y_{t-1}, y_t, X, t)\} \quad (2)$$

The most probable label sequence for an input X can be efficiently determined using the Viterbi algorithm.

$$Y^* = \arg\max_Y P_\Lambda(Y \mid X) \quad (3)$$

For the sequence labeling task like Chinese Word Segmentation, CRF and ME performed better than HMM. In addition to the advantages of the discriminative models, CRF optimizes parameters and decodes globally by taking state transition probabilities into account and consequently can avoid label bias problem. CRF is one of the most effective machine learning models for sequence labeling task.

We use CRF++ (version 0.55)[1] in this paper.

Table 1. Description of 6-tag tagset

Tags	Tag sequences for words of different lengths
S; B; B_2; B_3; M; E.	S; BE; BB_2E; $B\ B_2\ B_3E$; $B\ B_2\ B_3ME$; $B\ B_2\ B_3M\ldots ME$.

Table 2. Feature templates

Type	Feature	Description
Unigram	C_{-2} , C_{-1}, C_0 , C_1 , C_2	C_0 denotes the current character; $C_n(C_{-n})$ denotes the character n positions to the right (left) of the current character.
Bigram	$C_{-2}C_{-1}$, $C_{-1}C_0$, C_0C_1 , C_1C_2 , $C_{-1}C_1$	ditto
Punctuation	$IsPu(C_0)$	Current character is punctuation
Character Type	$K(C_{-2})K(C_{-1})K(C_0)K(C_1)K(C_2)$	Types of character: date, numeral, alphabet, others

2.2 Tag Set and Feature Template

We apply CRF to Chinese Word Segmentation by regarding it as a sequence labeling task. This is implemented by labeling the position of each Chinese character in word that the character belongs to. (Zhao et al., 2006) reported that a 6-tag set can achieve the best performance among the tag sets for Chinese Word Segmentation. Therefore we also use the same 6-tag set, whose definition is descripted in Table 1 in detail.

[1] http://crfpp.sourceforge.net/

Following the work of (Low et al., 2005), we adopt feature templates and a context of five characters for feature generation. The feature templates used in our model are shown in Table 2. Character Type in the bottom of Table 2 is the types of Chinese characters. Four types are defined in Table 3.We call the information as basic features.

3 N-gram Features

In this paper, n-gram refers to a sequence of n consecutive Chinese characters. A word can be considered as a stable sequence of characters. In a large enough corpus, words with some meanings will occur here and there repeatedly. This implies corresponding sequences of characters will be repeated in the corpus. Reversely, the sequences repeated in a large number of texts are more likely to be words. This is the basis on which n-gram features are used for word segmentation. In different domain, the ways of constructing words from characters are some different. When a target domain is

Table 3. Four types of Chinese characters

Type	Character set
date	年、月、日
numeral	1,2,3,4,5,6,7,8,9,0,一，二，三，四，五，六，七，八，九，零
alphabet	a,b,c,d,e,f,g,h,i,j,k,l,m,n,o,p,q,r,s,t,u,v,w,x,y,z,A,B,C,D,E,F,G,H,I,J,K,L,M,N,O,P,Q,R,S,T,U,V,W,X,Y,Z
others	Other characters

shortage of large annotated corpus, an approach proposed in this paper can be applied for automatic domain adaptation by exploiting the n-gram features from the raw corpus of the target domain.

In this work, we examined two kinds of n-gram features: n-gram frequency feature and AV feature. Our experimental results show that these simple statistical features are indeed effective in improving the ability of CWS system for domain adaptation.

3.1 N-gram Frequency Feature

We define n-gram frequency as the number of occurrences of n-gram in a corpus. The reason for considering this information is that the higher the frequency of n-gram, the greater the possibility of it being a word.

N-gram frequencies are extracted from raw corpus of target domain for (2=<n<=5) in this paper. In fact,we also tried n=6,but the results were not satisfied.Considering efficiency in computing, n-grams whose frequency values are less than five are filtered out. In order to alleviate the sparse data problem, we group all the frequency values into three sets: high-frequency (H), middle-frequency (M), and low-frequency (L) (Yiou Wang et al, 2011). The grouping way are defined as follows: if the frequency value of a n-gram is one of the top 5% of all the frequency values, the frequency value of this

n-gram is represented as H; if it is between top 5% and 20%, it is represented as M, otherwise it is represented as L. In this way, n-gram frequency lists are produced. We regard the n-gram as words in the following processing.

For CRF training and decoding, the features of current character are generate as follows. We retrieve the n-gram lists for candidate words that contain the current character. From a candidate word, a feature is generated in the form of "A-B", where A is the position of the current character in the candidate word and B is the frequecy of this candidate word. Then, the feature generated from each candidate word is concatenated with each other by "|" as one n-gram frequency feature. Note that the concatenating order follows the position orders of the current character in candidate words, i.e. B、B_2、B_3、M、E, for standardization.

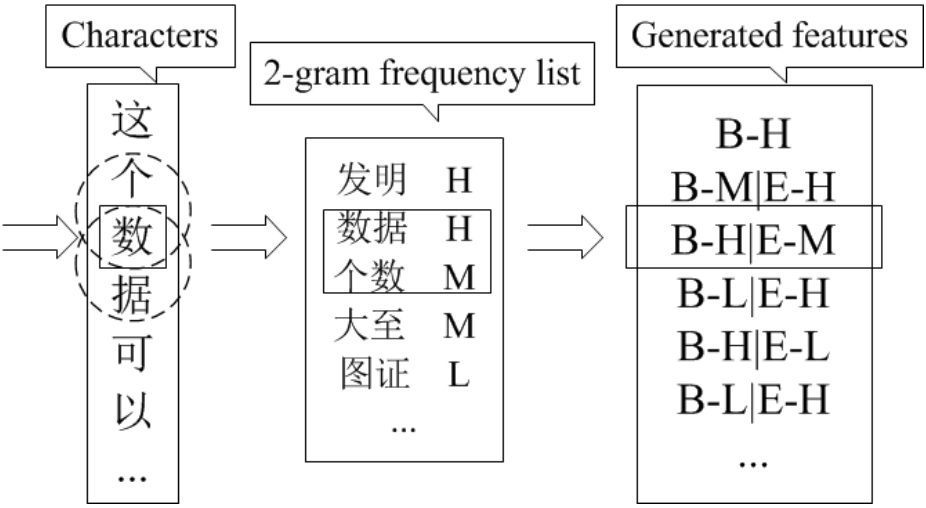

Fig. 1. Generation of 2-gram frequency feature

Fig.1 shows the generation process of 2-gram frequency feature for the sentence "这个数据可以…" (The data can …). The feature for the current character "数" (numeral), displayed in a square frame, is generated as follows. First, candidate words "个数" (number) and "数据" (data) are retrieved from the 2-gram frequency list. From the candidate word "个数" a feature "E-M" is generated because its frequency is "M" and "数" is the last character of the word. After that, another feature from the candidate word "数据" is generated as "B-H" because the frequency is "H" and "数" is the first character of the word. Finally, the 2-gram frequency feature is represented as "B-H|E-M".

3.2 N-gram AV Feature

AV (Accessor Variety) is a statistical standard used in (Feng et al., 2004) to determine whether a character sequence is a word when extracting words from Chinese raw texts.

(Hai Zhao and Chuyu Kit, 2007; Hai Zhao and Chuyu Kit, 2008) has explored an approach to extract AV global features from raw corpus for CRF learning. Following the work of (Luo and Huang, 2009), we focus on improving the way of generating features for CRF learning and on avoiding data sparseness at the same time.

Different from the n-gram frequency, the n-gram AV has a selection for frequency. The main idea of the AV is that if a character sequence appear in a variety of context, then the sequence is likely to be a word. The AV of a sequence *s* is defined as:

$$AV(s) = \min\{L_{av}(s), R_{av}(s)\} \tag{4}$$

Where $L_{av}(s)$ and $R_{av}(s)$ are defined, respectively, as the numbers of distinct predecessor and successor of *s*.

At first, AV feature of n-gram (2=<n<=5) are extracted from raw corpus. Then following the grouping way described in 2.1, the AV feature are grouped into three sets, high-frequency (H), middle-frequency (M), and low-frequency (L) and thus n-gram AV lists are produced. In generation of n-gram AV feature, candidate words containing current character will be retrieved out from the n-gram AV lists and then features of the candidate words are concatenated as the final n-gram AV feature for CRF training and decoding.

4 Case Study—Domain Adaptation of Chinese Word Segmentation to Patent Domain

In order to verify the proposed approach, we specify patent domain for CWS to adapt to. The raw corpus of the patent domain is taken from the Chinese part of the NTCIR-9[1] Chinese-English parallel patent description sentences. Such formed Chinese patent corpus consists of 1 million sentences. There are two phases in the domain adaptation implementation, construction of n-grim statistical information base and generating of n-gram features. In the first phase, n-gram frequency features and n-gram AV features are extracted from the corpus and n-gram statistical information base including n-gram frequency lists and n-gram AV lists is produced.In the second phase, n-gram features are generated for the sentences.

4.1 Construction of N-gram Statistical Information Base

According to the definitions of n-gram frequency feature and the n-gram AV feature described in section 2, we extracted character sequences of n-gram (2=<n<=5) from the Chinese patent corpus for construction of n-gram statistical information base. The overview of the construction of n-gram statistical information base is shown in Fig.2.

[1] http://research.nii.ac.jp/ntcir/ntcir-9/

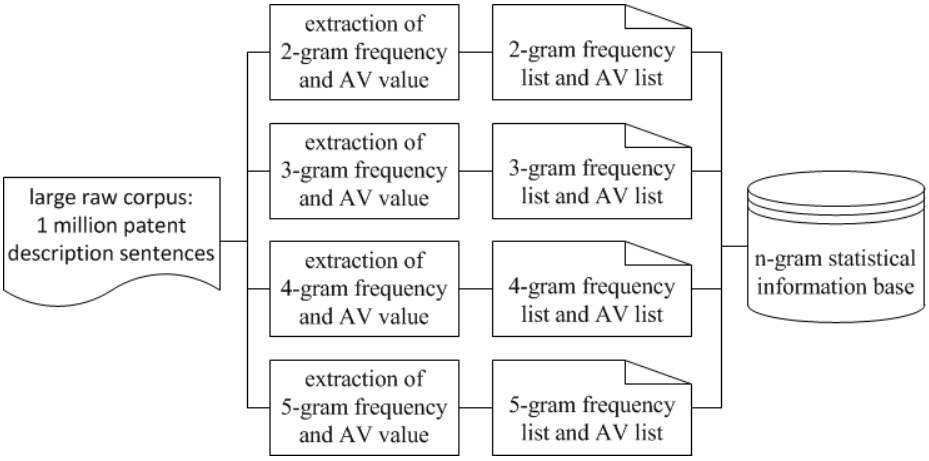

Fig. 2. Construction of n-gram statistical information base

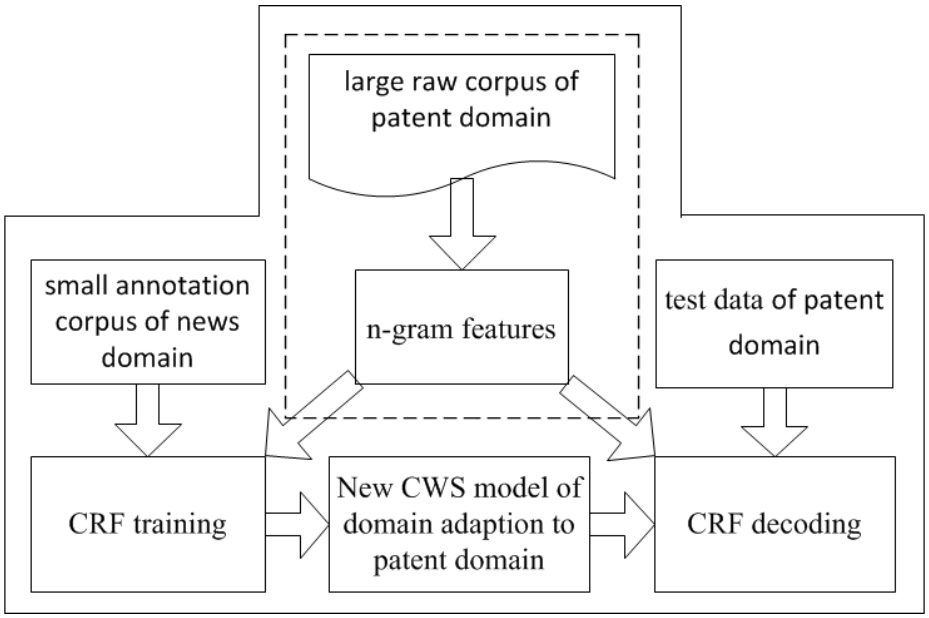

Fig. 3. Framework of domain adaptation of CWS system to patent domain

4.2 Generation of N-gram Features

In CRF training and decoding, n-gram features are generated for each character of sentences. The method of generating n-gram features has been described in 2.1 and 2.2, respectively. The framework of domain adaptation of CWS system to the patent

domain is shown in Fig.3. The part surrounded by the dotted line is the core component for the CWS system to adapt to the patent domain.

5 Experiments and Analyses

In order to verify the contribution of the n-gram features to domain adaptation of Chinese Word Segmentation, we evaluate the segmentation results of the new CWS system incorporated with the n-gram features of the patent domain and then compare the results with those of the baseline CWS system that uses only basic features.

5.1 Data

We used the Penn Chinese Treebank (CTB) as annotated corpus and defined data sets as follows: chapter 1-270, chapter 400-931 and chapter 1001-1151 for training data set; chapter 270-300 for test data sets. The proportion of unknown words is 3.47%. Since the data of the corpus is mainly from newswire, the domain of CTB may be regarded as news domain.

The unlabeled data of the patent domain is the Chinese patent corpus. The n-gram statistical information base is built from this raw corpus.

For test data of the patent domain, we randomly selected 300 sentences from the Chinese patent corpus and manually annotated word segmentations following the specification of Penn Chinese Treebank Project. As a result, 10636 Chinese words are obtained. By referring to the training data of CTB, we found that the proportion of unknown words in this patent test data is 22.4%.

5.2 Results and Analyses

We used recall (R), precision (P), and F_1 as evaluation metrics and also measured the recall on OOV (R_{OOV}) tokens and in-vocabulary (R_{IV}) tokens.

Table 4 shows the performances of the baseline system on the test data of CTB and the patent domain. The baseline system is developed on the training data of CTB by using the basic features. The proportions of unknown words in test data of CTB and the patent domain are respectively 3.47% and 22.4%.

Table 4 shows that the baseline system performed very well on CTB test data. This can be explained that the test data and the training data are from the same domain, i.e. the news domain. When test data changed to the patent domain, F_1 value of the baseline

Table 4. Performances of baseline CWS system on test data from different domains

CWS System	Source of test data	R	P	F_1	R_{OOV}	R_{IV}
Baseline	CTB(news domain)	98.02%	97.21%	97.62%	75.18%	98.85%
	NTCIR(patent domain)	86.05%	81.83%	83.89%	63.70%	92.51%

Table 5. Performances of different CWS systems on test data from the patent domain

CWS System	R	P	F_1	R_{OOV}	R_{IV}
Baseline	86.05%	81.83%	83.89%	63.70%	92.51%
+(a)n-gram frequency feature	87.95%	84.32%	86.09%	69.15%	93.37%
+(b)n-gram AV feature	88.29%	84.28%	86.24%	69.15%	93.82%
+(a)+(b)	88.31%	84.61%	86.42%	69.91%	93.63%

system greatly decreased to 83.89% from 97.62%. For a more detailed investigation, R_{OOV} decreased to 63.70% by 11.48% and R_{IV} decreased to 92.15% by 6.34%. The degree of decline on OOV is about twice that on in-vocabulary. These investigation results demonstrated the serious impacts brought to the performance of Chinese Word Segmentation by the changes of domain. The task of domain adaptation is therefore very important for Chinese Word Segmentation.

The performances of the CWS systems developed by using the proposed approach are shown in Table 5, where "a" refers to using n-gram frequency feature and "b" refers to using n-gram AV feature. The test data is the 300 sentences of the patent domain. For comparison, the result of the baseline system on the same test data is also shown in Table 5. The results show that both n-gram frequency feature and n-gram AV feature contributed to the improvement in performance from the view of each metric, and that the combination of (a) and (b) achieved further improvements.

Through the observation of Table 4 and Table 5, we may conclude as follows.

- The impact of interdisciplinary on Chinese Word Segmentation is very obvious, and the introduction of a large number of OOV will cause a serious decline in the performance of CWS system.
- The n-gram features including n-gram frequency feature and n-gram AV feature are very effective in each evaluation metric and the combination of them can achieve further improvement.
- In terms of F_1 measure, the improvement contributed by n-gram AV feature is greater than that contributed by n-gram frequency feature. In terms of recall, n-gram AV feature is more effective than n-gram frequency feature, while n-gram frequency feature is more effective than n-gram AV feature in terms of precision.
- N-gram features can effectively increase the recall of OOV in CWS system.

It is further observed that some scientific and technical terms were successfully recalled after n-gram features of the raw corpus have been added into the system. For instance, one manually annotated sentence "它/可用于/驱动/电光层/到/显示/数据…" (…which may be used to drive the electro-optic layer to a state in which display datum…) was segmented as "…/电光/层到/…" by the baseline system, while the new system obtained the correct result. For another manually annotated sentence: "而/不/使用/压电/元件/或/楔板/换能器" (…without piezoelectric element or wedge

transducer…), the baseline system segmented the word “换能器” into two parts, “换” and “能器” , while the new system avoided this erroneous segmentation. The word “电光层” and “换能器” are both physics terms which seldom occur in the news domain. The original intention of the paper is to explore n-gram features of the target domain corpus for a CWS system to be able to recognize the new words and the CWS system on the patent domain performed as we expected.

Through the above experimental results and analyses, we observed that n-gram features are effective for the CWS system to adapt to the patent domain from the news domain.

6 Conclusion

This paper proposes an approach in which n-gram features from large raw corpus are explored to realize domain adaptation for Chinese Word Segmentation. The n-gram features include n-gram frequency feature and AV feature. Our experiments on the patent domain show that the n-gram features are effective in domain adaptation of Chinese Word Segmentation. This approach can be easily implemented because the n-gram features can be extracted from a raw corpus of the target domain. Compared with an annotated corpus and a domain-specific dictionary, a raw corpus of the target domain is easily obtained with low cost.

In the future, we will conduct further study on how to explore more effective features from large raw corpus and how to incorporate them into statistical learning model for domain adaptation of Chinese Word Segmentation.

References

1. Zhang, M., Deng, Z., Che, W.: Combining Ststistical Model and Dictionary for Domain Adaptation of Chinese Word Segmentation. Journal of Chinese Information Processing (2012)
2. Lafferty, J., McCallum, A., Pereira, F.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: ICML (2001)
3. Zhao, H., Huang, C., Li, M., Lu, B.: Effective tag set selection in Chinese Word Segmentation via conditional random field modeling. In: PACLIC 2006, Wuhan, China, pp. 87–94 (2006)
4. Low, J.K., Ng, H.T., Guo, W.: A Maximum Entropy Approach to Chinese Word Segmentation. In: Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing (SIGHAN 2005), pp. 161–164 (2005)
5. Wang, Y., Kazama, J., Tsuruoka, Y., Chen, W., Zhang, Y., Torisawa, K.: Improving Chinese Word Segmentation and POS Tagging with Semi-supervised Methods Using Large Auto-Analyzed Data. In: Proceedings of the 5th IJCNLP, pp. 309–317 (2011)
6. Feng, H., Chen, K., Deng, X., Zheng, W.: Accessor variety criteria for Chinese word extraction. J. Computational Linguistics 30, 75–93 (2004)
7. Zhao, H., Kit, C.: Incorporating global information into supervised learning for Chinese Word Segmentation. In: PACLING 2007, Melbourne, Australia, pp. 66–74 (2007)

8. Zhao, H., Kit, C.: Unsupervised Segmentation Helps Supervised Learning of Character Tagging for Word Segmentation and Named Entity Recognition. In: Proceedings of the Six SIGHAN Workshop on Chinese Language Processing, Hyderabad, India, pp. 106–111 (2008)
9. Luo, Y., Huang, D.: Chinese Word Segmentation Based on the Marginal Probabilities Generated by CRFs. Journal of Chinese Information Processing 23, 3–8 (2009)
10. Xia, F.: The Segmentation Guidelines for the Penn Chinese Treebank (3.0) (2000)