A Kalman Filter Based Human-Computer Interactive Word Segmentation System for Ancient Chinese Texts

Tongfei Chen[1], Weimeng Zhu[1], Xueqiang Lv[3], Junfeng Hu[2, *]

[1] School of Electronics Engineering & Computer Science,
Peking University, Beijing, 100871, P. R. China
[2] Key Laboratory of Computational Linguistics, Ministry of Education,
Peking University, Beijing, 100871, P. R. China
{ctf,zwm,hujf}@pku.edu.cn
[3] Beijing Key Laboratory of Internet Culture and Digital Dissemination Research,
Beijing Information Science and Technology University, Beijing, 100101, P. R. China
lxq@bistu.edu.cn

Abstract. Previous research showed that Kalman filter based human-computer interaction Chinese word segmentation algorithm achieves an encouraging effect in reducing user interventions. This paper designs an improved statistical model for ancient Chinese texts, and integrates it with the Kalman filter based framework. An online interactive system is presented to segment ancient Chinese corpora. Experiments showed that this approach has advantage in processing domain-specific text without the support of dictionaries or annotated corpora. Our improved statistical model outperformed the baseline model by 30% in segmentation precision.

Keywords: Word Segmentation, Human-Computer Interactive System, Kalman Filter, Ancient Chinese Corpus Processing

1 Introduction

Since Chinese text is written without natural delimiters such as whitespaces, word segmentation is the essential first step in Chinese language processing [1]. Over the past two decades, various methods have been developed to address this issue [2–7]. Generally, supervised statistical learning methods are more robust in processing unrestricted texts than the traditional dictionary-based methods.

However, in some domain-specific applications, for example ancient Chinese text processing, there is neither enough annotated homogeneous corpora for training a reliable statistical model, nor a sufficient lexicon. Under these circumstances, unsupervised methods are preferred to utilize the linguistic knowledge derived from the raw corpus itself. Many researches also explores human-computer interactive segmentation process, enabling users to add expert knowl-

* To whom all correspondence should be addressed.

edge to the system [8, 9]. Since the criteria of word segmentation is sometimes dependent on users, interactive segmentation is reasonable [10].

Human-computer interactive approaches enables users to review and proofread the raw segmentation result produced by the statistical model. Zhu et al. proposed a Kalman filter based human-computer interactive learning model for segmenting Chinese texts depending upon neither lexicon nor any annotated corpus [11]. This approach enables users to observe and intervene the segmentation results, while the segmenter learns and adapts to these knowledge iteratively. At the end of this procedure a segmentation result that fully matches the demand of the user is returned.

This paper devises an improved model for ancient Chinese word segmentation, and uses the Kalman filter model by Zhu et al. to implement a practical system for human-computer ancient Chinese text processing.

The rest of this paper is organized as follows. The next section reviews related work. Our statistical model is introduced in Section 3. Section 4 briefly reviews the Kalman filter based approach. In Section 5, we presents the design of our practical segmentation system for ancient Chinese texts. In Section 6, the evaluation is presented, and the final section concludes this paper and discusses possible future work.

2 Related Work

Unsupervised word segmentation is generally based on some predefined criteria, such as *mutual information* (*mi*), to recognize a substring as a word. Sproat and Shih studied comprehensively in this direction using mutual information [12]. Many successive research applied mutual information with different ensemble methods [13, 14]. Sun et al. designed an algorithm based on the linear combination of *mi* and *difference of t-score* (*dts*)[15]. Other criteria like *description length gain* [16], *assessor variety* [17] and *branch entropy* [18] are also explored. Shi et al. adopted conditional random fields to generate a unified process for word segmentation and POS-tagging on pre-Qin ancient Chinest texts [19].

Any automatic segmentation has limitations and is far from fully matching the particular need of users. Thus human-computer interactive strategies are explored to allow users to bring their linguistic knowledge into the segmenter by intervening the segmentation process. Wang et al. developed a sentence-based human-computer interaction inductive learning method [8]. Feng et al. proposed a certainty-based active learning segmentation algorithm, which uses an EM (Expectation Maximization) algorithm to train an n-gram language model in an unsupervised learning framework [20]. Li and Chen further explored a candidate words based human-computer interactive segmentation strategy [21].

The Kalman filter [22] is an efficient recursive filter that estimates the internal state of a linear dynamic system from a series of noisy measurements. Recent researches have introduced Kalman filter model to promote user experience of Internet applications, by estimating click-through rate (CTR) of available articles (or other objects on web pages) in near real-time for news display systems

[23]. Zhu et al. applied Kalman filter model to learn and estimate user intentions in their human-computer interactive word segmentation framework [11].

3 Statistical Model

3.1 Baseline Model

Sun et al. proposed *difference of t-score* (*dts*) [3] as a useful complement to *mutual information* (*mi*) and designed a compound statistical measure based on the linear combination of *mi* and *dts*, named *md* [15].

$$mi^*(x,y) = \frac{mi(x,y) - \mu_{mi}}{\sigma_{mi}} \ , \tag{1}$$

$$dts^*(x,y) = \frac{dts(x,y) - \mu_{dts}}{\sigma_{dts}} \ , \tag{2}$$

$$md(x,y) = mi^*(x,y) + \lambda \cdot dts^*(x,y) \ , \tag{3}$$

λ is set as an empirical value 0.6 in Sun's paper; $mi(x,y)$ is the normalized mutual information of any given bigram xy, and $dts(x,y)$ is the normalized difference of t-score of bigram xy. Given any bigram xy, in terms of $md(x,y)$ and a threshold Θ, whether this bigram be combined or seperated can be determined — when $md(x,y)$ is greater than Θ, the bigram xy is marked as *combined*; otherwise, it is marked as *separated*.

There exists a possible optimization scheme when a local minumum or maximum of *md* appears [3]. Consider a Chinese character string $wxyz$. If $md(x,y) > md(w,x)$ and $md(x,y) > md(y,z)$, $md(x,y)$ is called a *local maximum*. *Local minima* follow a similar definition. It can be seen that even a $md(x,y)$ of a local maximum does not reach the threshold Θ, xy may still be combined, while if $md(x,y)$ of a local minimum is greater than Θ, xy is more likely to be separated despite its *md* value. To reflect this kind of tendency, we increase the *md* values at local maxima by a constant s, and decrease the *md* values at local minima by s.

This statistical model will be used as a baseline model in further discussions.

3.2 Improved Statistical Model

For bigrams with a smaller number of occurrences, the statistical measure of mutual information (*mi*) is not reliable. We define a weight for mutual information, i.e.

$$w(x,y) = \log_2 \left(f(x,y) + 1 \right) \ , \tag{4}$$

where $f(x,y)$ is the frequency of bigram xy in the corpus.

Additionally, some proper nouns (e.g. names of people or places) tends to occur only in several adjacent paragraphs or chapters. This rendered the mutual

information of these words low, resulting in these words to be judged as *separate*. If a bigram recurs frequently, i.e. clumps in context, it is more likely to be a content-bearing word [24, 25]. We define a *bigram recurrence* to measure this tendency. Define *bigram recurrence* as

$$br(x, y) = \log_l f_l(x, y) \ , \tag{5}$$

where l is length of context chosen; and f_l is the frequency of bigram xy in context window of length l.

Combine the baseline model and the measures we define above, we define

$$A(x, y) = \lambda_i w(x, y) mi^*(x, y) + \lambda_t dts^*(x, y) + \lambda_r br^*(x, y) \ , \tag{6}$$

where br^* denotes the normalized version of br, and λ_i, λ_t and λ_r are coefficients. If $A(x, y)$ is greater than a threshold Θ, xy is judged as *combined*; otherwise, it is judged as *separated*. We call this function as the A feature. A feature combined global measurements such as mi, as well as local measurement br which is dependent on contexts.

These parameters are trained using an annotated version of *Annals of the Five Emperors, Records of the Great Historian* (《史记 • 五帝本纪》). These values are chosen as

$$\lambda_i = 0.43, \lambda_t = 1.0, \lambda_r = 0.37, \Theta = 1.0 \ . \tag{7}$$

The local minima and maxima optimization described in Section 3.1 is also exploited in this improved model.

3.3 Structural Words Optimization

In Classical Chinese, some structural words seldom form words with other characters. These characters are judged as single-character words directly in our model. Structural words chosen in this paper includes the following characters:

而, 何, 乎, 乃, 其, 且, 若, 所, 为, 焉, 也, 以, 因, 于, 与, 则, 者, 之, 弗, 莫, 不, 哉, 矣, 又, 已

For example, in character sequence xyz, if y belongs to the structural word set above, the values $A(x, y)$ and $A(y, z)$ are all set to be below the threshold value Θ so that bigram xy and yz are both judged as *separated*.

4 Kalman Filter Model

Zhu et al. developed a human-computer interactive learning word segmentation algorithm using Kalman filters [11]. This model is equipped with a Kalman filter to make it learn and estimate user intentions from the interventions (which may contain noise) for each bigram. The linguistic knowledge is gradually accumulated from the process of user interactions, and eventually, a segmentation result

that fully matches the need of the user (or with an accurate rate of 100% by manual judgement) is returned within limited times of interventions. A basic assumption is that each bigram (of different characters) is independent, i.e., if the state of one bigram is modified, states of other bigrams are not affected.

In this section, we adapt the Kalman filter model by Zhu et al. to the improved statistical model described in Section 3. For simplicity, we focus on a specific bigram xy.

4.1 Process State

A time step is defined as a manual judgement to the segmentation result of a bigram. Since that the system is viewed as a human interaction process, it can also be mapped to a time series process. Given a bigram xy, we assume the statistical measure A in the corpus follows a stable Gaussian distribution $N(\mu, \sigma^2)$, where μ and σ^2 are the expectation and variance of A respectively. We define x_t is the *system state* of the the A feature of a specific bigram at time t. the state of time $t+1$ is estimated upon time t :

$$\hat{x}_{t+1|t} = \hat{x}_t + w_t \ , \tag{8}$$

where w_t represents the uncertainty (i.e. noise) of this prediction at time t which follows a normal distribution. This distribution is formulated as

$$w_t \sim N\left(0, Q_t\right) \ , \tag{9}$$

where Q_t is the autocovariance of the bigram at time t. It can be calculated as

$$Q_t = \mathrm{E}\left[\left(\hat{x}_{t-2} - \mu_{t-2}\right)\left(\hat{x}_{t-1} - \mu_{t-1}\right)\right] \ , \tag{10}$$

where

$$\mu_t = \mathrm{E}[\hat{x}_t] \text{ for each } t \ . \tag{11}$$

Kalman filter also predicts the variance of the state change, which is

$$P_{t+1|t} = P_t + Q_t \ , \tag{12}$$

where P_t represents the estimation of the state variance at time t.

4.2 Measurements and States Update

Since the system is human-computer interactive, a measurement system that maps manual judgements to a continuous space of A feature is introduced. Apparently, some uncertainty (i.e. observation noise) is inevitable, and we assume that it follows a normal distribution. To guarantee that the mapped values corresponds to the manual judgments, we take a high confidence interval (for example 99%). The system measurements z_t of the true state x_t is assumed to be generated according to

$$z_t = x_t + v_t \ , \tag{13}$$

where v_t is the uncertainty of observations which is assumed to be a Gaussian white noise R_t. After the observation, The Kalman filter will update the prediction of next state using a Kalman gain [22]:

$$K_t = \frac{P_{t+1|t}}{P_{t+1|t} + R_{t+1}} \ . \tag{14}$$

The state prediction of time $t+1$ is updated as

$$\hat{x}_{t+1} = \hat{x}_{t+1|t} + K_t(z_{t+1} - \hat{x}_{t+1|t}) \ , \tag{15}$$

and the updated state variance is estimated as

$$P_{t+1} = (1 - K_t)P_{t+1|t} \ . \tag{16}$$

Then, this updated state can be used to segment next time this bigram appears in the corpus.

5 Human-Computer Interactive System

In our practical system[1], the user first load the raw corpus into the system. The system will first segment the whole text using the improved statistical model elaborated in Section 3. The interface of the system after loading the raw corpus is shown in Figure 1.

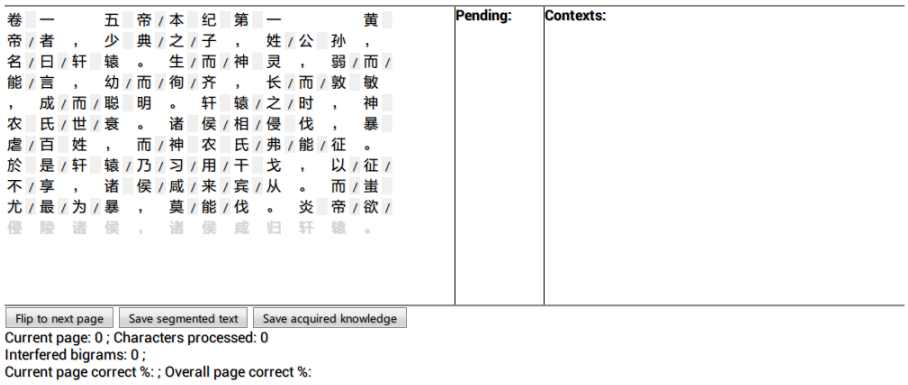

Fig. 1. Snapshot of our system just after loading *Annals of the Five Emperors, Records of the Grand Historian* (《史记 • 五帝本纪》).

Every bigram's status — whether *combined* or *separated* — can be modified by a click on the symbol between the two characters of a bigram. Modified

[1] The system can be found at `http://klcl.pku.edu.cn/clr/ccsegweb/kalman_segmenter.aspx`.

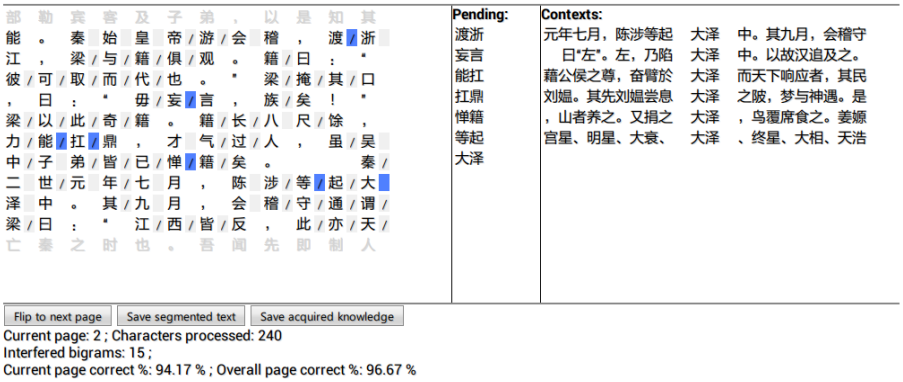

Fig. 2. Snapshot of our system while segmenting *Annals of Xiang Yu, Records of the Grand Historian* (《史记 • 项羽本纪》).

bigrams will be marked with a different color. These clicks act as user input to the system. When the user flips to the next page, pending user interventions will be applied to Kalman filters of these intervened bigrams. Different contexts of the current modified bigram is shown on the right panel of the interface, from which users can check the meaning of this bigram under different contexts.

These features of the system is shown in Figure 2.

During the segmentation process, the system keeps track of the changes the user made, hence it is able to produce better segmentation results as the human-computer interaction progresses. Users can save the current segmentation result and the states of the Kalman filters at any time.

6 Experiments

In this section, we conducted several experiments to evaluate our model. Firstly, we verified the improvement after introducing our new statistical model. Secondly, we verified the effectiveness of Kalman filter model in reducing human effort. The ancient Chinese corpus used for experiments are chapters from *Records of the Grand Historian* (《史记》) and *History of Song* (《宋史》).

As there is no standard specification for ancient Chinese segmentation, we used experts to segment *Annals of Xiang Yu, Records of the Grand Historian* (《史记 • 项羽本纪》, abbreviated as *Xiang Yu*, approximately 11000 characters) and a part of *Annals of Taizu I, History of Song* (《宋史 • 本纪第一 • 太祖一》, abbreviated as *Taizu I*, approximately 2000 characters) as test corpora.

6.1 Improved Statistical Model

In this part, we verified the effectiveness of our improved statistical model without the Kalman filter based human-computer interaction process. Thus, our

improved statistical model acts as an automatic segmenter without the human-computer interaction process. The baseline model used for comparison is the model by Sun et al. [15].

To evaluate the performance of these models, we use the precision and recall rate, as well as the *accuracy of segmentation* (abbreviated as 'Accuracy' in this paper) described by Sun et al. in [3]. It is defined as

$$\text{Accuracy}[\%] = \frac{\#\text{ of locations being correctly marked}}{\#\text{ of locations in corpus}} \times 100\% \,. \tag{17}$$

Corpus for tests are the aforementioned *Xiang Yu* and *Taizu I*. The results are shown in the following two tables.

Table 1. Different measures for *Xiang Yu*

	Accuracy	Precision	Recall
Sun's Approach	78.18%	54.71%	59.21%
Our model	**90.79%**	**86.94%**	**80.55%**

Table 2. Different measures for *Taizu I*

	Accuracy	Precision	Recall
Sun's Approach	74.57%	45.60%	55.45%
Our model	**88.35%**	**75.71%**	**66.44%**

From these tables above, it can be seen that our model significantly outperformed Sun's model because Sun's model is more suitable to handle contemporary Chinese texts, while our model is optimized on ancient Chinese texts. Our model achieved an improvement of more than 30% in segmentation precision; and achieved an improvement of about 13% $\sim$ 14% in terms of accuracy of segmentation.

Since our model is trained using a text excerpt from *Records of the Grand Historian* (《史记》), *Xiang Yu* is a homogeneous corpus, while *Taizu I* is a heterogeneous corpus. On homogeneous text such as *Xiang Yu*, our model yields a satisfactory result, achieving 86.94% in precision and 80.55% in recall. On heterogeneous text *Taizu I*, a text written more than 1000 years after *Records of the Grand Historian* (《史记》) is completed, the result was still acceptable.

6.2 Kalman Filter Model

In this part, we simulated the human-computer interaction by using the correct segmentation text as input to the model so as to evaluate the performance of the Kalman filter model. We adopted the *binary prediction rate* (BPR) described by Zhu et al. [11] to quantify the conformity of the prediction in the model with

user intention,

$$\text{BPR}[\%] = \frac{\#\text{ of correct predictions}}{\#\text{ of all predictions}} \times 100\% \,. \tag{18}$$

Two models were compared in this section. One is the approach discussed in Section 6.1, i.e. our improved statistical model without the human-computer interaction process is abbreviated as *Without learning*), and the other is the Kalman filter integrated approach discussed in Section 4 (abbreviated as *Kalman approach*).

The result of the experiment is shown in Table 6.2. Corpora used is the same as the previous section.

Table 3. The BPR[%] of different approaches.

Corpus	*Xiang Yu*	*Taizu I*
Without learning	90.79%	88.35%
Kalman approach	**92.38%**	**88.86%**

From this experiment, it can be seen that on homogeneous text such as *Xiang Yu*, Kalman filter based human-computer interactive model outperformed the baseline statistical value by about 1.6%. On *Taizu I*, the improvement was insignificant because the text is rather short (approximately 2000 characters).

7 Conclusions and Future Work

Previous research showed that Kalman filter based human-computer interaction Chinese word segmentation algorithm achieves an encouraging effect in reducing user interventions. This paper designs an improved statistical model for ancient Chinese texts, and integrates it to the Kalman filter based framework, resulting in a practical system. Experiments revealed that our improved statistical model significantly outperforms the baseline model, and the Kalman filter approach achieves an notable improvement in reducing human efforts.

Our future work will focus on establishing an interactive bootstrapping segmentation system with an accumulating dictionary.

Acknowledgments. This work is partially supported by Open Project Program of the National Laboratory of Pattern Recognition (NLPR) and the Opening Project of Beijing Key Laboratory of Internet Culture and Digital Dissemination Research (ICDD201102).

References

1. Liang, N.Y.: CDWS: An Automatic Word Segmentation System for Written Chinese Texts. Journal of Chinese Information Processing, 2(2): 44-52(1987) (in Chinese)

2. Nie, J.Y., Jin, W., Hannan M.L.: A Hybrid Approach to Unknown Word Detection and Segmentation of Chinese. In: Proceedings of the International Conference on Chinese Computing, pp. 326-335 (1994)
3. Sun, M., Shen, D., Tsou, B.K.: Chinese Word Segmentation Without Using Lexicon and Hand-Crafted Training Data. In: COLING/ACL 1998, pp. 1265-1271 (1998)
4. Luo, X., Sun, M., Tsou, B.K.: Covering Ambiguity Resolution in Chinese Word Segmentation Based on Contextual Information. In: COLING 2002, pp. 1-7 (2002)
5. Zhang, H.P., Liu, Q., Cheng, X.Q., Yu, H.K.: Chinese Lexical Analysis Using Hierarchical Hidden Markov Model. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, pp. 63-70 (2003)
6. Peng, F., Feng, F., McCallum, A.: Chinese Segmentation and New Word Detection Using Conditional Random Fields. In: COLING 2004, pp. 23-27 (2004)
7. Goldwater, S., Griffiths, T.L., Johnson, M.: Contextual Dependencies in Unsupervised Word Segmentation. In: COLING/ACL 2006, pp. 673-680 (2006)
8. Wang, Z., Araki, K., Tochinai, K.: A Word Segmentation Method with Dynamic Adapting to Text Using Inductive Learning. In: Proceedings of the First SIGHAN Workshop on Chinese Language Processing, pp. 1-5 (2002)
9. Li, M., Gao, J., Huang, C., Li, J.: Unsupervised Training for Overlapping Ambiguity Resolution in Chinese Word Segmentation. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, pp. 1-7 (2003)
10. Sproat, R., Gale, W., Shih, C., Change, N.: A Stochastic Finite-State Word-Segmentation Algorithm for Chinese. Computation Linguistics, 22(3): 377-404 (1996)
11. Zhu, W., Sun, N., Zou, X., Hu, J.: The Application of Kalman Filter Based Human-Computer Learning Model to Chinese Word Segmentation. In: Computational Linguistics and Intelligent Text Processing, pp. 218-230 (2013)
12. Sproat, R., Shih, C.: A Statistical Method for Finding Word Boundaries in Chinese Text. Computer Processing of Chinese and Oriental Languages, pp. 336-351 (1990)
13. Chien, L.F.: Pat-Tree-Based Keyword Extraction for Chinese Information Retrieval. In: ACM SIGIR Forum, pp. 50-58 (1997)
14. Yamamoto, M., Kenneth, C.W.: Using Suffix Arrays to Compute Term Frequency and Document Frequency for All Substrings in a Corpus. Computer Linguistics, 27(1): 1-30 (2001)
15. Sun, M., Xiao, M., Tsou, B.K.: Chinese Word Segmentation without Using Dictionary Based on Unsupervised Learning Strategy. Chinese Journal of Computers, 27(6): 736-742 (2004) (in Chinese)
16. Kit, C., Wilks, Y.: Unsupervised Learning of Word Boundary with Description Length Gain. In: Proceedings of the CoNLL99 ACL Workshop, pp. 1-6 (1999)
17. Feng, H., Chen, K., Deng, X., Zheng, W.: Accessor Variety Criteria for Chinese Word Extraction. Computation Linguistics, 30(1): 75-93 (2004)
18. Jin, Z., Tanaka-Ishii, K.: Unsupervised Segmentation of Chinese Text by Use of Branching Entropy. In: COLING/ACL 2006, pp. 428-435 (2006)
19. Shi, M., Li, B., Chen, X.: CRF Based Research on a Unified Approach to Word Segmentation and POS Tagging for Pre-Qin Chinese. Journal of Chinees Information Processing, 24(2): 39-45 (2010) (in Chinese)
20. Feng, C., Chen, Z., Huang, H., Guan, Z.: Active Learning in Chinese Word Segmentation Based on Multigram Language Model. Journal of Chinese Information Processing, 20(1): 50-58 (2006) (in Chinese)
21. Li, B., Chen, X.: A Human-Computer Interaction Word Segmentation Method Adapting to Chinese Unknown Texts. Journal of Chinese Information Processing, 21(3): 92-98 (2007) (in Chinese)

22. Kalman, R.E.: A New Approach to Linear Filtering and Prediction Problems. Journal of Basic Engineering, 82(1): 35-45 (1960)
23. Agarwal, D., Chen, B.C., Elango, P., Motgi, N., Park, S.T., Ramakrishnan, R., Roy, S., Zachariah, J.: Online Models for Content Optimization. In: Proceedings of NIPS 2008, pp. 17-24 (2008)
24. Liu, Z., Sun, M.: Web-Based Automatic Detection for IT New Terms. Proceedings of the 9th China National Conference on Computational Linguistics: 515-521 (2007)
25. Bookstein, A., Klein, S. T., Raita T.: Clumping Properties of Content-bearing Words. Journal of the American Society for Information Science 49(2) : 102-114 (1998)