Example-Based Sense Tagging of Running Chinese Text

Xiang Tong
Chang-ning Huang
Cheng-ming Guo

Computer Science Department
Tsinghua University
Beijing, CHINA

Phone: +86-1-2594895
Fax: +86-1-2562768

ABSTRACT

This paper describes a sense tagging technique for the automatic sense tagging of running Chinese text. The system takes as input running Chinese text, and outputs sense disambiguated text. Whereas previous work (Yarowsky, 1992; Gale, et al., 1992, 1993) relies heavily on the role of statistics, the present system makes use of Machine Readable/Tractable Dictionaries (Wilks, et al., 1990; Guo, in press) and an example-based reasoning technique (Nagao, 1984; Sumita, et al., 1990) to treat novel words, compound words, and phrases found in the input text.

Key words: sense tagging

1. Introduction

If the 1980's were characterized by the surge of efforts on Machine Readable/Tractable Dictionary (MRD/MTD) research, the 1990's would be a time of massive efforts on constructing annotated text corpora. Properly annotated text corpora could form, at least, the bases for the following:

a. the core of commercial information systems;
b. the kernel engine of 'Cognitive Agents';
c. the essentials of systems vital to national security.

Sense tagging of large text corpora has been on the back-burner for too

long. The preparation of large annotated text corpora, especially those with word sense disambiguated, has always been brushed aside for some piteous 'smart' approaches to prevail. However, it is just this kind of hopeless cleverness that handicapped the speedy growth of the language enterprise. Fortunately, more and more researchers have come to realize the importance, as well as the necessity, of being earnest in annotating large text corpora of all major languages.

The present discussion presents a system for the automatic sense tagging of running Chinese text —— a necessary mechanism for the construction of annotated 'Monitor Corpora' (Sinclare, 1991) that do not degrade over time. The system takes as input running Chinese text, and outputs sense disambiguated text. Whereas previous work (Yarowsky, 1992; Gale, et al., 1992, 1993) relies heavily on the role of statistics, the present system makes use of Machine Readable/Tractable Dictionaries (Wilks, et al., 1990; Guo, in press) and an example-based reasoning technique (Nagao, 1984; Sumita, et al., 1990) to treat novel words, compound words, and phrases found in the input text. The focus of this discussion is on the example-based reasoning technique. The examples that support the tagging operation come from the system MTD.

The sense tagging system assigns a unique number for every Chinese characters occurred in the text. In most cases, the senses tagged are word senses. This is due to the fact that most Chinese characters are words. For example, '打' (beat) has 26 senses. '鼓' (drum) has 6 senses. The phrase '打鼓' (beat drums) becomes '打_B02 鼓_A01' after sense tagging. However, not all Chinese characters are words. Sometimes they are bound morphemes. In these cases, the senses tagged are the meanings of the morphemes as given in the dictionary. For example, '阿' as in '阿爸', '阿爹' is tagged 'A01', which is the number of '阿' as given in the MTD when '阿' is used as a prefix, i. e., a bound morpheme.

2. Overview of the Sense-Tagging System

The sense-tagger under discussion represents partial results of some three years of continued efforts on the part of Tsinghua University, Beijing, China to build systems for the processing of general, unrestricted running Chinese texts. The system was implemented in 'C', and currently runs on the Sun Workstation at the National AI Laboratory in the University.

2.1. Resources

The sense-tagging module uses two MRDs and one MTD. The first MRD,

for the sake of discussion, say MRD_1,is '现代汉语通用字典' (Fu, 1987). It contains about 6,000 one-syllable words, e. g. , '打' (beat), '鼓' (drum), and 43,000 compound words and phrases, e. g. '打鼓' (beating drums). Each word has one or more word senses. For example, '打' (beat) has 26 senses and '鼓' (drum) 6. Note that capital letters in the numbers tagged indicate homographs, and the Arabic numbers the sense number under the homograph. The content of the word '打' (beat) is given as following:

打_A01: 量词,十二个叫一打

打_B01: 殴打, 攻打
打_B02: 用手或器具撞击物体
打_B03: 做, 从事
打_B04: 表示身体上的某些动作
打_B05: 定出, 计算
打_B06: 放射, 发出
打_B07: 涂抹, 画, 印
打_B08: 除去
打_B09: 捉(禽兽等)
打_B10: 揭, 凿开
打_B11: 采取某种方式
打_B12: 做某种运动或游戏
打_B13: 捆
打_B14: 建造, 修筑
打_B15: 发生与人交涉的行为
打_B16: 买
打_B17: 付给或领取(证件等)
打_B18: 编织
打_B19: 制造(器物、食品等)
打_B20: 搅拌
打_B21: 从
打_B22: 用割、砍等动作来收集
打_B23: 酌取
打_B24: 举, 提
打_B25: 器皿、蛋类等因撞击而破碎

The second MRD, for the sake of discussion, say MRD_2, is the Chinese thesaurus '同义词词林' (Mei, 1983) with about 70,000 entries. It has a 3-level categorization system. At Level 1, the dictionary has 12 major categories. At Level 2, the 12 major categories split into 94 subcategories. At the lowest level, Level 3, the dictionary has altogether 1,428 subcategories. Under the current numbering system, the capital letter indicates major categories, the lower-case letter subcategories, and the Arabic numbers the numbering under the two superordinate categories. For example, 'Bp13' refers to one of the categories that the word '鼓' (drum) falls into. B is a first level category, p is a second level subcategory, 13 is the numbering of the subcategory under Bp. Partial list of the numbering of some categories is given as follows:

鼓	Bp13 Fa01 Fa30 Ic05	雇	Hc26
谷	Be04 Bho5	瓜	Bh07
灌	Fa23 Hd16	馆	Dm05
锣鼓	Bp13	箩筐	Bp10
裸露	Id03	落成	Iel4

The MTD was constructed from MRD_1. It has 43,000 annotated compound words and phrases. Word phrases like '打鼓' (beating drums) are disambiguated in the MTD with word sense numbers tagged to both '打' (beat) and '鼓' (drum), e.g. '打_B02 鼓_A01'. The numbers tagged are based on the numbering system as used in MRD_1. For those compounds that have component whose meaning is not related to the resultant compound, the Arabic numbers in the component's tag is '00' (e.g., 白_A00 痴_A01, 玲_A00 珑_A00). Much of the work in constructing the MTD was done by machine, but supplemented by handcoding. The following gives a partial list of the contents of the MTD :

打_B01	倒_A03	打_B01	手_A02	挨_B01	打_B01
打_B02	鼓_A02	打_B02	火_A01	打_B02	门_A01
打_B03	劫_A01	打_B03	捞_A01	打_B05	量_A02

2.2. Three-step Sense-tagging Procedure

Step 1 : Segmenting the input text into words, compound words and phrases

The word segmentation module is a much simplied version of a more complicated segmentation program developed at the Laboratory. It looks forward through each sentence for maximum match of character strings as recorded in the MTD. The tagging of most known phrases is done with the help of the MTD. '打鼓' would be an example in question. The involved operation is simple, i.e., 'match to access'. When an input segment matches an entry in the MTD, the tagged form of the matched segment replaces the input segment in the sentence.

Step 2 : Example-based sense tagging of one-syllable words

The system uses an example-based sense-tagging algorithm for the disambiguation of one-syllable words, which are not listed in system MTD. The detail of the algorithm is described in Section 3.

Step 3 : Default sense tagging of untagged one-syllable words from Step 2

A default sense number is assigned to each and every one syllable word untagged from Step 2. The default sense numbers are determined on the basis of frequency of occurrence data.

3. Example-Based Sense-Tagging

Chinese words build to form compound words. In 94.7% of the time, the meaning of the resultant compounds is related to the contributing meanings of the component words (Zhang, 1986, p. 87). The compound words and phrases in the MTD contain implicit syntactic information for purpose of example-based reasoning about the senses of Chinese words in context.

For example, if '打 锣鼓' (beat gongs and drums) is in the input text and the sense of '打' (beat) cannot be determined. In order to disambiguate the word sense of '打' (beat), the system looks through the MTD for every compound word and phrase beginning with '打' (beat) and decides that the phrases '打_B02 鼓_A01' (beat drums) is an appropriate example to reason about the word '打' (beat) as found in '打 锣鼓' (beat gongs and drums), since '鼓' (drums) and '锣 鼓' (gongs and drums) are in the same lowest category 'Bp13' in MRD_2. The system then assigns the tag 'B02', which belongs to '打' (beat) in '打_B02 鼓_A01' (beat drums), to '打' (beat) in '打 锣鼓' (beat gongs and drums).

Formally, when $S_1\ S_2 \cdots\ S_n$ represent input segments from 1 to n, W represents an untagged segment, and the immediate context of W is represented by $L_{range} \cdots\ L_2\ L_1\ W\ R_1\ R_2 \cdots\ R_{range}$, where L stands for 'Left', R stands for 'Right', and *range* equals 5, we have the following:

$$S_1\ \ S_2\ \ \cdots\ \ S_n \qquad (a)$$

where $S_k (k=1, \cdots, n)$ is a word, compound word or phrase

$$L_{range}\ \ \cdots\ \ L_2\ \ L_1\ \ W\ \ R_1\ \ R_2\ \ \cdots\ \ R_{range} \qquad (b)$$

where L_i, $R_i (i=1, \cdots, range)$ is a word, compound word or phrase

In the forward reasoning process, assuming that ($W\ R_i$) is a possible compound word or phrase, for all entries in MTD beginning with W which is in the form ($W\_tag\ Item$), the system computes the relatedness of the two words or phrases ($W\ R_i$) and ($W\_tag\ Item$), where '*Item*' may be an annotated word, compound word, phrase, or just a meaningless Chinese character string. The concept distance of R_i and *Item* is computed to determine the relatedness of the two compound words/phrases. Hence,

Concept_Distance(R_i, *Item*) =

0	if R_i and *Item* are in the same lowest category in MRD_2
1	if R_i and *Item* are in the adjacent categories in MRD_2
100	all other cases

Relatedness(($W\ R_i$), ($W\_tag\ Item$)) =

2 if Concept_Distance(R_i, *Item*) = 0

1 if Concept_Distance(R_i, *Item*) = 1

0 if Concept_Distance(R_i, *Item*) = 100

For every pair of (W R_i) ($i=1, \cdots, range$) and (W_tag *Item*) in the MTD, the pair that has the greatest non-zero relatedness measure is determined and the W in (b) above is substituted by the W_tag in the determined pair.

The reasoning process works similarly in both directions of W, i. e., forward to R_{range} and backward to L_{range}. When the process proceeds forward, the system looks for entries beginning with W. On the other hand, when the process works backwards to the left of W, the system looks for annotated entries in the MTD ending with W.

The examples are given as following:

(1) 因此,利用 单倍体 植株 培育 *新* 品种,可以 明显 地 缩短 育种 年限。

The word '新' (new) has six senses. The annotated phrase '新_A01 型_A01' is found in the MTD. The system calculates the conceptual distance between '型' and '品种' among others. Since '型' and '品种' are found to be in the same lowest subcategory 'Dd06', the conceptual distance between them is 0. The system then assigns the tag 'A01', which belongs to '新' as in '新_A01 型_A01', to '新' in the above sentence.

(2) 对 人民 负责 ,*受* 人民 监督 。

The word '受' (receive, suffer) has six senses. The annotated phrase '受_A02 审_A02' is found in the MTD. The system calculates the conceptual distance between '审' and '监督' among others. Since '审' and '监督' are found to be in the adjacent lowest subcategories, i. e., 'Hc18' and 'Hc19' respectively, the conceptual distance between them is 1. The system then assigns the tag 'A02', which belongs to '受' as in '受_A02 审_A02', to '受' in the above sentence.

(3) 国家 保护 公民 的 合法 的 收入 、储蓄 、房屋 和 其他 合法 财产 的 所有 *权* 。

The word '权' (right, power) has seven senses. The annotated phrase '财_A01 权_A01' is found in the MTD. The system calculates the conceptual distance between '财' and '财产' among others. Since '财' and '财产' are found to be in the same lowest subcategory 'Dj03', the conceptual distance be-

tween them is 0. The system then assigns the tag 'A01', which belongs to '权' as in '财_A01 权_A01', to '权' in the above sentence.

(4) 使 人口 的 增长 同 经济 和 社会 发展 计划 *相* 适应。

The word '相' (each other) has four senses. The annotated phrase '相_A01 称_A01' is found in the MTD. The system calculates the conceptual distance between '称' and '适应' among others. Since '称' and '适应' are found to be in the same lowest subcategory 'Jc01', the conceptual distance between them is 0. The system then assigns the tag 'A01', which belongs to '相' as in '相_A01 称_A01', to '相' in the above sentence.

4. Evaluation

The input Chinese texts that the present system works on are news release texts from the official Chinese Xinhua News Agency. No preprocessing of these news release texts is required.

The performance of the present sense-tagger is encouraging. The hit rate of correct sense tagging can run as high as 95%. The lowest hit rate ever recorded was 70%. The appendix gives a sample text which is the output of our system. The hit rate of correct sense tagging of this sample is 93.79%. Essentially, the hit rate of correct sense tagging performed by the system is a function of the coverage of the system MTD and MRDs.

5. Limitations and Future Work

a. The system makes errors when the segmentation of the input texts is less than correct. The performance of the current sense tagger can be improved if more sophisticated segmentation method is adopted.

b. Although the reasoning process takes advantage of collocational information within the phrase in which the untagged segment is a part, there is no guarantee that the phrase does not have multiple meanings. When such cases occur, the result of the reasoning is subject to chance.

c. The example-based sense tagging method works quite well with content words, but for function words it often makes faulty guesses. This is partly due to the fact that function words are less sensitive to context. The current system assigns a default sense number for most function words. However, for those words which can both be a function word and a content word, the system often makes errors. This kind of errors decreases when the system preprocesses the input texts with a stochastic Chinese grammatical tagger like the one developed at Tsinghua University (Bai, et al., 1992).

6. Conclusion

In this paper we presented a relatively simple but effective method for the sense tagging of running Chinese texts. The system takes advantage of the collocation information within the annotated compound words or phrases in the system MTD. Considering that annotated Chinese texts constitute very useful resources for Chinese language processing, especially in generating frequency of occurrence/co-occurrence data, general and special purpose concordances and the data for the derivation of a natural set of semantic primitive for the Chinese language, the current sense-tagging system looks promising. The room for progress is to be found in the further improvement of the system resources and the refinement of the reasoning algorithm.

References

Bai, S-H, Xia, Y., Huang, C-L. (1992) Research on Chinese grammatical tagging method for Chinese corpus. In: Chen, Z-X (Ed.), *Development in machine translation*. Dianzi Gongye Publishing House: Beijing. pp. 408-418.

Fu, X-L., (1987) '现代汉语通用字典', Waiyu Jiaoxue Yu Yanjiu PublishingHouse: Beijing.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992) Work on statistical methods for word sense disambiguation. *Working Notes for AAAI Fall Symposium on Probabilistic Approaches to Natural Language*. pp. 54-60.

Gale, W. A., Church, K. W., and Yarowsky, D. (1993) A method fordisambiguating word senses in a large corpus. To appear in: *Computers and Humanities*.

Guo, C-M (in press) Machine Tractable Dictionaries: Design and Construction. Ablex: Norwood, NJ.

Mei, J-Zh. (1983), '同义词词林', Shanghai Cishu Publishing House: Shanghai.

Nagao, M. (1984) A framework of a mechanical translation between Japanese and English by analogy example. In: A. Elithorn, R. Benerji, (Eds), *Artificial and Human Intelligence*. Elsevier: Amsterdam.

Sinclare, J. (1991) Monitor corpora. *Corpus, Concordance, Collocation*. Oxford University Press. pp. 24-26

Sumita, E., Iida, H. and Kohyama, H. (1990) Translating with examples: a newapproach to machine translation. *The Third International Conference on Theoretical and Methodological Issues in Machine Translation of Natural Language*. Austin, Texas.

Wilks, Y., Fass, D., Guo, C-M, McDonald, J., Plate, T., and Slator, B. M. (1990)Providing machine tractable dictionary tools. *Journal of Machine Translation*. 5, 2, pp. 99-151.

Yarowsky, D. (1992) Word-sense disambiguation using statistical models of Roget's categories trained on large corpora. *COLING*-92

Zhang, W. (1986) *Character Meanings and Word Meanings*. China Wu Zi Publishing House: Beijing.

Appendix

Samples from Annotated Text

我_A02国_A01 胃_A01癌_A01 研_A01究_A01 接_A03近_A02 世_A01界_A02 水_A01平_A01

新华社_NAM 沈阳_LOC 5_NUM 月_A01e 3_NUM 日_A04e 电_A01e （_PUN 记_A01者_A01 徐延安_NAM ）_PUN 有_A01关_A01 部_A02门_A07 提_B01供_A01 情_A02况_A01 说_B01# ，_PUN 我_A02国_A01 对_A12# 胃_A01癌_A01 的_A01# 研_A01究_A01 和_A06# 诊_A01治_A02 已_A01# 逐_A01步_A03 接_A03近_A02 世_A01界_A02 先_A01进_A01 水_A01平_A01 。_PUN

据_B04e 对_A12# 全_A01国_A01 5 0_NUM 多_A05e 家_A01e 主_A01要_B01 医_A03疗_A01 科_A02研_A01 单_B00位_A01 的_A01# 统_A01计_A02 ，_PUN 近_A01e 十_A01e 几_B01e 年_A05e 里_A01# 累_B01计_A02 发_A00现_A01 早_A01期_B02 胃_A01癌_A01 1 4 0 0_NUM 多_A00e 例_A01e ，_PUN 治_A02疗_A01 效_A01果_A02 良_A01好_A01 。_PUN 其_B02中_A03 中国_LOC 医_A02科_A01 大_A01学_A01 全_A01部_A01 胃_A01癌_A01 的_A01# 5_NUM 年_A02e 生_A03存_A03 率_A01e 已_A01# 达_A02e 百_A02分_A02 之_A01# 五_A01e 十_A01e 八_A02e 点_A16e 五_A01# 。_PUN

我_A02国_A01 胃_A01癌_A01 发_A02病_A01率_A01 和_A06# 死_A01亡_A02 率_A01e 居_A03e 各_B01种_B02 癌_A01症_A01 前_A01列_A02 。_PUN 1 9 8 4_NUM 年_A01# ，_PUN 中国_LOC 医_A02科_A01 大_A01学_A01 与_B01# 日本_LOC 同_A01行_A01 开_A05展_A04 合_B01作_B01 ，_PUN 分_A01别_A03 在_A02# 我_A02国_A01 东北地区_LOC 和_A06# 日本_LOC 北海道_LOC 设_A01立_A01 了_A01# 胃_A01癌_A01 研_A01究_A01 基_A02地_B01 ，_PUN 试_A01图_A03 通_A01过_B01 对_A12# 当_A01地_B01 居_A01民_A01 的_A01# 胃_A01# 肠_A00e 状_A01况_A01 进_A01行_B02 观_A01察_A01 和_A06# 对_A01比_A01 ，_PUN 弄_A01# 清_A02e 导_A01致_A01 胃_A01癌_A01 的_A01# 饮_A02食_A01 和_A06# 环_A02境_A03 因_A01素_A01 。_PUN

近_A01e 几_B01e 年_A02来_A02 ，_PUN 中国_LOC 医_A02科_A01 大_A01学_A01 一_A01直_A07 把_A08# 胃_A01癌_A01 的_A01# 早_A01期_B02 发_A00现_A01 、_PUN 早_A01期_B02 诊_A01断_A02 、_PUN 早_A01期_B02 治_A02疗_A01 以_A02及_A04 胃_A01癌_A01 的_A01# 生_A10物_A01 学_A02e 行_B05为_A01 作_B02为_A02 研_A01究_A01 方_A06向_A02 。_PUN 早_A01e 在_A02# 1 9 7 2_NUM 年_A01# ，_PUN 中国_LOC 医_A02科_A01 大_A01学_A01 就_A09# 发_A00现_A01 国_A01内_B01 首_A03e 例_A01e 早_A01期_B02 胃_A01癌_A01 患_A02者_A01 ，_PUN 迄_A01今_A01 为_A02止_A05 他_A02们_A01

的_A01# 胃_A01癌_A01 早_A01期_B02 发_A00现_A01 率_A01# 已_A01#
超_A01过_B02 百_A02分_A02 之_A01# 二_B01# 十_A01e 。_PUN
此_A01外_A01 ，_PUN 在_A02# 对_A12# 中_A03# 晚_A02期_B02
胃_A01癌_A01 的_A01# 治_A02疗_A01 过_B01程_A02 中_A03# ，_PUN
他_A02们_A01 区_B03别_A03 胃_A01癌_A01 的_A01# 不_A01# 同_A01#
恶_B02性_A01 度_A01# ，_PUN 开_A05展_A04 合_B02理_A01 的_A01#
胃_A01癌_A01 扩_A01大_A01 根_A04治_A01 手_A01术_A01 及_A04#
其_B02他_A01 辅_A01助_A01 疗_A01e 法_A02e ，_PUN 治_A02疗_A01
水_A01平_A01 得_A01到_A01 明_A04显_A02 提_B04高_A02 。_PUN

专_A01e 家_A01e 认_A02为_A00 ，_PUN 随_A01着_C01 胃_A01癌_A01
研_A01究_A01 水_A01平_A01 的_A01# 进_A01e 一_A01e 步_A01e
提_B04高_A02 和_A06# 人_A01民_A01 生_A09活_A03 状_A01况_A01
的_A01# 不_A01断_A01 改_A01善_A02 ，_PUN 胃_A01癌_A01 将_A02#
因_A04# 发_A00现_A01 期_B02e 提_B06前_A01 而_A02# 越_A01e
来_A07# 越_A01e 容_A00易_A01 医_A03治_A02 ，_PUN 胃_A01癌_A01
发_A02病_A01率_A01 也_A01# 将_A02# 逐_A01渐_B01 下_A02降_A01
。_PUN
（_PUN 完_A02# ）_PUN

'@': indicator of example-based sense tagging
'#': indicator of default sense tagging