An Analytical Study of Transformational Tagging for Chinese Text

Helen Meng* and Chun Wah Ip
Human-Computer Communications Laboratory
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Email:{hmmeng, cwip}@se.cuhk.edu.hk
Fax: (852)2603-5505
*Please send all related correspondences to this author

6 Aug 1999

Topic Areas: (o), (s)

Abstract

This work is our initial attempt in using the transformation-based error-driven learning (TEL) procedure for tagging Chinese text. TEL has previously been shown to be effective in POS tagging for English [Brill 1995]. TEL provides several attractions: (i) automation for tagging, (ii) induction of interpretable rules, (iii) learning aimed at error-reduction. Our experimental corpus consist of over 70,000 words of Chinese text, divided into disjoint training and test sets of a 9:1 ratio. With an unknown word/tag proportion of 13%, we achieved overall tagging accuracies of 94.56% (training) and 86.87% (testing).

1. Introduction

Part of speech tagging is an important linguistic problem which has garnered much research interest and effort over the years. Automatic part of speech (POS) taggers are particularly attractive for providing syntactic information applicable to speech recognition and understanding, information retrieval, machine translation and other applications. A myriad of techniques have previously been used for automatic POS tagging, ranging from rule-based to data-driven approaches. The former tends to be hand-annotated by linguistic experts, while the latter includes stochastic n-grams, HMMs, neural networks, trigger-pair predictions, genetic algorithms, etc. [Bai et al., 1992][Kupiec 1992][Lua 1996][Black 1998].[1] Rule-based approaches are linguistically well-motivated, but expert handcrafting is often an expensive and tedious process. Data-driven approaches attempts to ameliorate the tedium by capturing relevant linguistic constraints from a corpus of annotated data. However, the linguistic constraints captured are encoded in a large body probabilities and statistics, which do not lend themselves well for exploratory linguistic analysis.

Brill [Brill 1995] had previously proposed an alternative technique of transformation-based error-driven learning for automatic POS tagging in English. This approach combines the merits of rule-based and data-driven techniques in an elegant manner. The algorithm may be initialized randomly or with some linguistically-motivated specifications. Machine learning then proceeds with an annotated corpus, and with the objective of maximizing tagging accuracy. Such learning produces a compact rule set, which encodes the contextual and lexical constraints for tagging, and are easily interpretable by humans for studying the linguistic cues for POS tagging.

This work explores the use of transformation-based error-driven learning (TEL) for POS tagging (or transformational tagging) of Chinese text. The Chinese language presents a unique set of characteristics for the tagging algorithm, which include:

(i) The ideographic (character-based) nature of Chinese, in contrast to the alphabetic nature of English. Chinese text consists of strings of characters separated by punctuation marks. A Chinese word may consists of a single character, or multiple characters with no delimiters between words. Hence, Chinese text need to be *segmented* to form sequences of words. For a given string of characters, there may exist multiple legitimate segmentations. Different segmentations lead to different word sequences and hence different sequences of POS tags. In this work, our task is simplified by using a pre-segmented corpus.

[1] Informative citiations are many, those included here are by no means exhaustive.

(ii) Aside from the ambiguity caused by multiple segmentations, a given word may have multiple possible POS assignments. For example, 白a 馬ng and 馬nf 步芳npf[2], where 馬 is a common noun in the former and a last name of person in the latter.

(iii) The lexical structure of the Chinese word is very different compared to English. Inflectional forms are minimal, while morphology and word derivations abide to a different set of rules. A word may inherit the syntax and semantics of (some of) its compositional characters, for example, 紅 means *red* (a noun or an adjective), 色 means *color* (a noun), and 紅色 together means *the color red* (a noun) or simply *red* (an adjective). Alternatively, a word may take on totally different characteristics of its own, e.g. 東 means *east* (a noun or an adjective), 西 means *west* (a noun or an adjective), and 東西 together means *thing* (a noun). Yet another case is where the compositional characters of a word do not form independent lexical entries in isolation, e.g. the characters in 彷彿 (a verb) do not occur individually.

This work examines the utility of transformational tagging for Chinese text. We are especially interested in the linguistic rules induced automatically by TEL for individual Chinese words, as well as across a sequence of multiple words. Chinese linguistic structures may be observed in such rules, including grammar, morphology and word derivations. TEL is applicable not only to in-vocabulary words, it is also designed to handle the occurrences of unknown words in corpora.

2. Corpus and Tags

This work is based on the pre-segmented and hand-tagged corpus from Tsinghua University [Bai et al., 1992]. This news corpus is derived from the People's Daily (Renmin Ribao) in the year 1993. Altogether there are 112 articles and 71,804 words of running text, distributed across five domains: computer, military, science, technology and general news. Unique vocabulary entries exceed 9,000. Information about the entire corpus is tabulated in Table 1, and the word count in the table refers to

[2] These are word/tag pairs extracted from our corpora

the length of running text. Table 2 displays some example sentences from each domain, which shows the word/tag pairs for each sentence. In this work, we only tackle the tagging problem – our tagger learns from pre-segmented and tagged training sets, and tests on a pre-segmented test sets.

Domain	No. of Articles	# of Words (train)	# of Words (test)
Computing	10	5,479	509
Military	23	12,243	1,787
Science	20	12,922	1,391
Technology	20	11,383	1,228
News	39	22,358	2,505

Table 1: Distribution of Training and Testing Sets from the Tsinghua news corpus.

Domain	Example sentences
Computing	我們/rn 根據/p 這/rn 一/mx 漢化/vg 策略/ng 對/p DECnet-DOS/xch 進行/vgv 了/utl 分析/vgo 。/。
Military	反/vgn 機降/ng 成為/vgn 戰鬥/vgo 的/usde 重要內容/ng 。/。
Science	17世紀英國的/usde 醫學家/ng 哈維/npf ,/,
Technology	工作/ng 模式/ng 是/vy 當前科技/ng 情報/ng 體制/ng 改革/vg 中/f 的/usde 又/d 一/mx 熱點/ng 。/。
News	共同體/ng 將/va 參與/vgn 德國統/vg 問題/ng 的/usde 討論/vg ,/,

Table 2. Example sentences from our corpus.

The original tag set found in the Tsinghua corpus consists of 108 unique labels. These were exhaustively enumerated in [Lua 1996]. Out of this set, 25 are for punctuation, and the remaining ones draw fine distinctions for Chinese parts of speech. As an example, nouns are divided into 5 types: **nf** (last name), **npf** (name of person), **npu** (name of organization), **npr** (other proper nouns) and **ng** (common noun). We added an extra tag, **nvg**, to represent words which can either be a noun or a verb, such as 運動 (exercise) or 表示 (express/expression). The reason is as follows: in the original tagged corpus, there are words like 運動 which are tagged as general verbs, e.g.

雖然/cf 經歷/vgn 了/utl 各/rn 種/qnk 運動/vg 變化/vg ，/，

where the tags are: cf (連詞前段), vgn (帶體賓動詞), rn (體詞性代詞), qnk (種類量詞) and vg (一般動詞)

In this context, however, 運動 seems to play a role more similar to a noun, which motivated the design of the **nvg**[3] because 運動 in this case is not suitable to tag as word.

One may wonder whether the full tag set is necessary for Chinese POS tagging.[4] A preliminary investigation of our entire corpus reveals that approximately 100 tags occurred, with the most frequent one being **ng** (common noun), which occurred about 25.5% of the time. The most frequent 18 tags (which include a few punctuation tags) covers 80% of our running text corpus, while the most frequent 32 tags already covers 90%. Nevertheless, we proceeded with the full set of 109.

The ambiguities found in the Tsinghua corpus is 1.88 tags per word. (Please see Figure 1) Over 40% of the vocabulary can be tagged multiple ways. Out of this, the maximum number of tags per a word is 8. Table 3 lists the 8 POS tags of the word (表示) and their contexts.

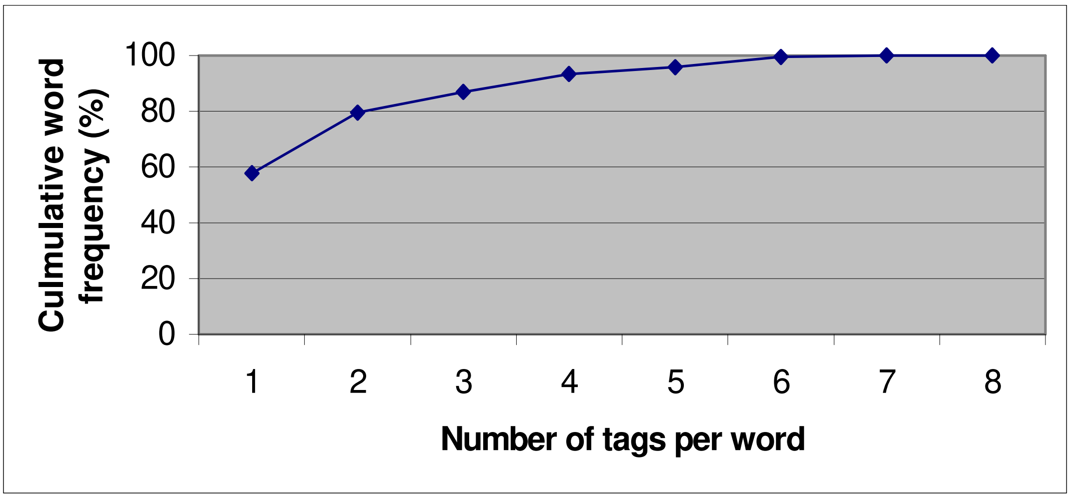

Figure 1: Cumulative distribution of words with single to multiple POS tags

[3] The idea of using nvg tags is attributed to Dr. Wenjie Li.
[4] We have found tag sets of approximately 50 entries of fewer in other literature.

Tag	Example Sentences
vg (一般動詞	這/rn 種/qnk 表示/vg 法/ng 與/p Ｊ Ａ/xch Ｅ 類似/a 。/。
vgo (不帶賓動詞	文獻/ng 和/cpw 查詢/vg 都/d 用/vgn 一/mx 組/qnc 正文/ng 詞/ng 向量/ng 表示/vgo ，/，
vgn (帶體賓動詞	我們/rn 采用/vgn Ｓ ｅ /xch Ｃ 表示/vgn 單/b 字節/ng 的/usde ＡＳＣ/xch 字符/ng ，/，
vgv (帶動賓動詞	即 /xch 人力/ng 表示/vgv 將/d 寫/p 此/rn 繼續/v 做/vgv 出/vc 努力/vgo 。/。
vga (帶形賓動詞	“ 是/vy ”/” 可以/a 表示/vga 一樣/a 。/。
vgs (帶小句賓動詞	情報/nd 表示/vgs ：/：
ng (普通名詞	情報/ng 檢索/vg 系統/ng 所/ng 儲存/ng 的/usde 是/vy 文獻/ng 的/usde 某/rn 種/qnk 表示/ng ，/，
nvg (動名詞	一/mx 篇/qni 文獻/ng 的/usde 表示/nvg 中/f 所/ussu 使用/vgn 的/usde 標引詞/ng 的/usde 個數/ng …

Table 3. Example sentences of the word “表示” from our corpus.

3. Transformational Tagging

The algorithm is presented in detail in [Brill 1995]. The tagger addresses its problem at both the *lexical* and *contextual* levels. Here we will provide a procedural sketch.

3.1 Notations

For the sake of simplicity, we will adopt the following notations in describing our work:

- C^{d}_{type} , denotes a corpus C belonging to a specific domain d, and of a particular *type* - *training, testing, lexical* or *contextual*. The type is related to the transformational tagging procedure, and will be explained later.
- $T_i(C^{d}_{type})$, denotes a *tagged* corpus C . The variable i may adopt the instances *ref* (for the set of reference tags), *start* (for the tags resulting from the initialization of the tagger) or *final* (for the tags resulting from the final stage of the tagger, having applied all tagging rules). Details will be explained later. An example of a tagged sentence is:

  17世紀英國的/usde 醫學家/ng 哈維/nf ，/，

- $U(C^{d}_{type})$, denotes a *untagged* corpus C. A procedure may be applied to strip off all the tags, resulting in 17世紀英國的醫生哈維

  *f*rom the previous example.

- R^{d}_{type} , denotes a set of rules *R*. Rules may be of the type *lex* (lexical rules) or *context* (contextual rules). Example rules include:[5]

  Lexical rule : goodleft vgn 135.820116353036
  Contextual rule : vgn vgo NEXT1OR2TAG STAART

  The associated explanation is in the following section.

- L_{type} , denotes a lexicon, which may be of type *lex* (the lexicon for training lexical rules only), or *all* (the lexicon containing all words in the training corpus).

3.2 Corpus Utilization

Figure 2 shows how the corpus is utilized. The entire corpus is first divided into a training set (90% of the size) and a test set (10%). The training set is in turn divided into two halves. One half is used to train *lexical* rules -- these are rules applied in order to predict the tag of a word based on the intra-word characteristics. The other half is used to train *contextual* rules -- these are rules applied to tag a word based on its neighboring word contexts.

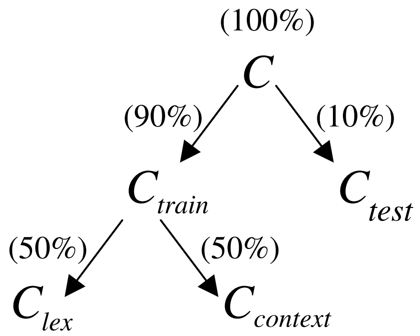

Figure 2. Corpus utilization in a particular domain.

3.3 Transformational-based Error Driven Learning

Learning takes place in two phases. Lexical rules are learnt first, and are used during the subsequent learning of contextual rules.

---

[5] The associated explanation is in the following section.

3.3.1 Lexical rules

These are used to tag unknown words. Learning lexical rules requires three word lists:

(i) A list of all the words occurring in the untagged training corpus $U(C_{train})$, sorted by decreasing frequency of occurrences. The word list is used to find the most common prefixes and suffixes.

(ii) A list of triplets [word tag count] derived from $T_{ref}(C_{lex})$, e.g.

是	vy	365
和	cpw	358
在	pzai	339

The words with more than one tags will get different entries in the list. Besides the triplets [和cpw 358], the list also contains three more triplets, [和p 13], [和cpc 1] and [和cpw 1] . The count of the triplet is the frequency of the word tag pair in the tagged training corpus. The tagged words are used to calculate the weights of possible tags for a given word.

(iii) A list of word bigrams found in the untagged training corpus, $U(C_{train})$, e.g.

是利用

都採用

心還

The bigrams list is used to calculate the weight of the tags to the preceding/following word .

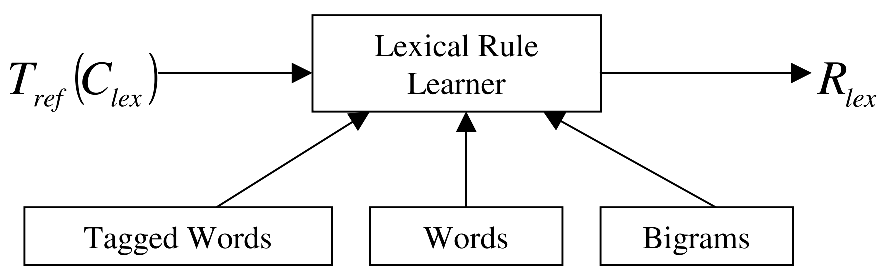

Figure 3. Learning Lexical Rules.

The learning process begins by giving the unknown word an initial tag. Such initialization can be done in a number of ways: The unknown word may be assigned **unk**, to denote its out-of-vocabulary nature. Alternatively, since unknown words are often common nouns, we may assign them with the tag **ng** upon initialization. In addition, we may utilization simple prior knowledge, e.g. assign **xch** (tag for non-chinese word) if English letters are encounted, or **mx** (tag for numbers used in measurements).

Lexical rules are learnt according to some prescribed templates, so that they can utilize prefixes, suffixes, constitutent characters and bigram relationships to infer an appropriate tag for an unknown word. Some example templates include:

- **{x w fgoodright/fgoodleft y n}**, i.e. given the word in focus **wc** currently tagged as **x**, should the word **w** occur to its right/left, change its tag from **x** to **y**. A close variant of this template is {w goodright/goodleft y n}, which does not constrain the current tag of the word in focus. **n** reflects the relative frequency of rule application in the training set. Here is the equation for calculating n.

$$n = \sum_{j=1}^{W} N\{word_j, tag_k\} - N\{word_j, tag_i\}$$

where ***W*** is the number of words in the training set, tag_k is target tag to be changed, tag_i is current tag.

$$N\{word_j, tag_k\} = \frac{word_j, tag_k}{\sum_{i=k}^{T} word_j, tag_i} \quad ,$$

where $\boldsymbol{word_j}$ is a word in the training set, $\boldsymbol{tag_k}$ is a tag for the $\boldsymbol{word_j}$, $\boldsymbol{T}$ is the number of tags for the $\boldsymbol{word_j}$, $\boldsymbol{word_j\ tag_k}$ is the number of frequency for the pair $\boldsymbol{word_j\ tag_k}$ in the training set.

Example of rule application:

Rule: { ng 李 fgoodright npf 11}

Sentence: 年/ng 過/vgn 半百/mx 的/usde 煉鐵廠/nc 老/a 工人/ng 李/nf 傳杰/npf

Here 傳杰 is a unknown word, and the tagger assigns it with **ng** upon initialization. However, seeing the last name 李 towards its left (i.e. 李 is to the *right* of our current word) invokes the specified rule. 傳杰 is then correctly transformed as a **npf** (name of a person).

- **{x z fchar y n},** i.e. given the word in focus **wc** currently tagged as **x**, should the character **z** occur in the word, change its tag from x to **y**. A close variant is {z char y n} which does not constrain the current tag of the word in focus. Example of rule application:

  Rule: mx 年 fchar t 46

  Sentence: 1957年/t 7月/t 到/p 1958年/t 12月/t

  The unknown word 1957年 will be tagged as **mx** (number for measurements) upon initialization. This invokes the specified rule to change to the correct tag **t** (tag for time).

- **{x a fhassuf/fhaspref p y n},** i.e. given the word in focus **wc** currently tagged as **x**, should it contain the **p** characters in its prefix or suffix **a**, change its tag from **x** to **y**. A close variant of this template is {a hassuf/haspref p y n}. Example of rule application:

  Rule: 委員會 hassuf 3 npu 5

  Sentence: 聯合國常規軍備委員會/npu 通過/vgn 決議/ng , / , The unknown word 聯合國常規軍備委員會 is initialized as **ng**. Owing to the occurrence of suffix 委員會 the tag will be changed to **npu** (name of organization).

Therefore it can be seen that the lexical rules automatically learnt during this stage offers insight as to the lexical nature of the words, interpreted with the use of prefixes, suffixes, constitutent characters as well as bigram information.

3.3.2 Contextual rules

The use of lexicons and lexical rules ensure that each and every word in the text is initialized with a tag. Contextual rules need to be learnt in order to correct any possible errors in the initialization. Hence these rules should be effective in disambiguating among the multiple tag assignments for a given word, using across-word contextual information. The learning process for contextual rules is depicted in Figure 4.

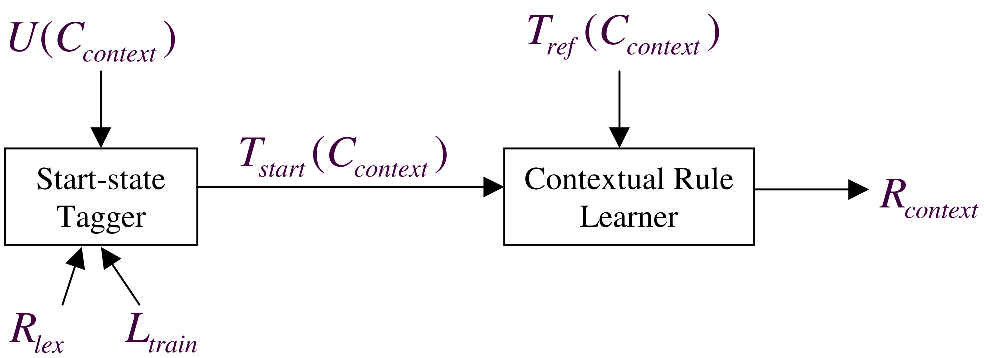

Figure 4. Flow chart showing the process of learning contextual rules

The untagged corpus for learning contextual rules is first processed by the start-state tagger. This tagger references the training lexicon, L_{train}, to assign the most frequent tag to each of the words. Unknown words are tagged by applying the lexical rules. These procedures produce a set of start-state tags T_{start} $(C_{context})$ for the corpus. These are then compared with the reference tags, T_{ref} $(C_{context})$, in order to proceed with error-driven learning, which finally produces the set of contextual rules $R_{context}$. Error-driven learning of the contextual rules also follow a set of templates, which considers the across-word context in a seven-word window - between one to three words/tags to the left and right of the current word (word in focus). Examples of the templates include:

- **{x y next1or2tag staart},** i.e. given that the current word **wc** is tagged as **x**, change the tag to **y** if the following one or two tags is the start/end of sentence symbol (**staart**).

  Example of rule application:

  Rule: usde y next1or2tag staart

  Sentence: 全/a 過程/ng 中/f 是/vy 可/va 變/vgo 的/y 。/。

  的 is most commonly tagged as **usde**, and is initialized by the start-state tagger thusly. Application of our rule corrects the assignment from usde to y (語氣詞)

- **{x y prevwd w},** i.e. given the current word wc is tagged as **x**, change the tag to **y** if the previous word is **w**. Example application:

  Rule: vv f prevwd 年

  Sentence: 寫/vi 過去15/mx 年/ng 來/f 的/usde 最/d 大/a 跌/vg 幅/ng 。/。

  The most frequent tag of 來 is **vv**, which becomes the initial assignment of the start-state tagger. However, the application of the rule corrects it to **f** (方位詞)

  During the learning process, the start-state tags are compared with the reference tags for each sentence in $C_{context}$. Rules for error correction are proposed according to the templates. The proposed rule which maximally reduces the number of errors is adopted in the *ordered* transformational rule set. The adopted transformation is then applied to the entire training corpus, from left to right, and the transformation is invoked only after all matching contexts in the training set are identified. This constitutes one iteration in learning. Iteration continues until no proposed rules can reduce the minimum count of tagging errors. This minimum count threshold is therefore an experimental parameter.

The difference between the templates of lexical rules and contextual rules is that lexical rules only consider the lexical information of the words (such as prefix, suffix and characters in the word) and neighbouring words. For contextual rules, the considerations are contextual information (such as the previous/following tag of current word), lexical information (such as the previous/following words of current word) and combination of lexical and contextual information (such as the previous/following word and previous/following tag together).

4. Experiments

Our experiments are based on disjoint training and test sets, with a 9:1 divide. Each corpus domain is processed individually. We have also combined all the articles for all domains to form a large corpus (71,804 words). This is also divided into training and test sets of the same proportion, and used for experimentation. Figure 5 displays a couple of example sentences.

UNIX/xch Pacific/xch 公司/ng 與/p AT&T/xch 是/vy 什麼 關係/ng ? (UNIX) (Pacific) (company) (and) (AT & T) (is) (what) (relationship)
它/rn 主要 是/vy 幹/vgn 什麼 的/usde ? (It) (mainly) (is) (doing) (what)

Figure 5. Examples from the training set, with both segmentation and tagging included. We also include a pseudo English translation in parentheses.

Since the training and test sets are disjoint, we see the occurrences of both *unknown words* as well as *unknown tags* in the test set. An "unknown tag" refers to the tagging of a (known) word in the test set, but the word/tag combination never appeared in the training set. For example, the single-character word 幹 was only seen with the tag **vgn** in the training set. However, it occurred in the test set with the tag **vgv**. Our tagger is bound to make mistakes with cases of unknown tags. The proportion of unknown words and unknown tags range from 8.95% to 33.20% across our domains. Details are shown in Table 4.

Domain / Proportion(%)	Computing	Military	Science	Technology	News	Total

Unknown Words	29.08	13.26	22.14	7.33	15.85	10.00
Unknown Tags	4.13	4.31	3.31	1.63	3.07	2.99
Unknown Words & Tags	33.2	17.57	25.45	8.96	18.92	12.99

Table 4. Distribution of unknown words and unknown tags in the test sets across domains.

4.1 Lexical Tag Initialization

As mentioned in the previous section, there are multiple schemes for assigning the initial tag to an unknown lexical entity. We can either assign it as **unk** (unknown), **ng** (common noun, most frequently occurring tag for unknown words), or according to our *initial assignment rule*, which incorporates a small amount of prior knowledge:

If the word contains an English letter (A-Z / a-z), tag it as ***xch*** *(non-chinese word)*

Else tag as ***ng*** *(common noun).*

Results comparing the three schemes are shown in Figure 6.

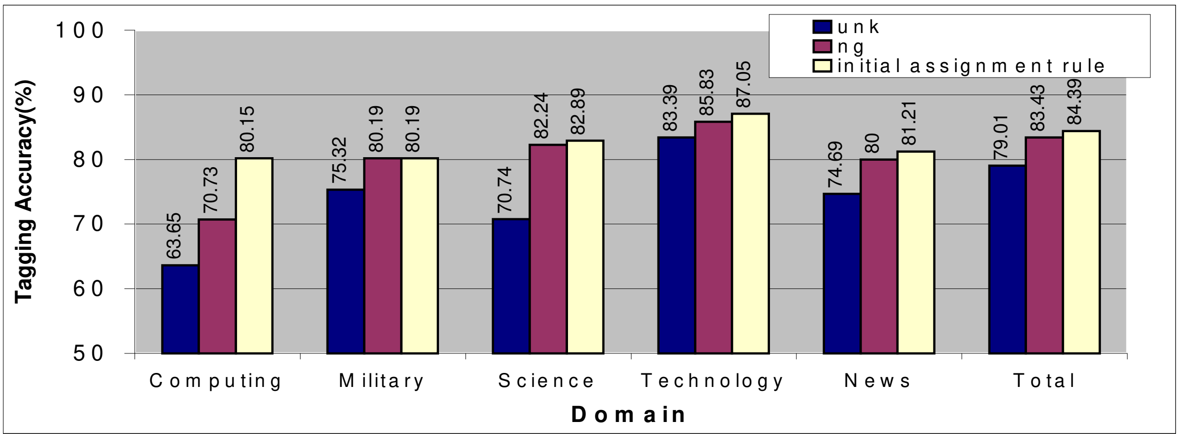

Figure 6. Test-set tagging accuracies (%) for the three different initial assignment schemes across the various domains.

Our initial assignment rule fares better than the straightforward **unk** or **ng** assignments. Hence we have decided to adopt it for our experiments.

4.2 Contribution of Lexical and Contextual Rules

Having acquired the initial stage assignments T_0, we proceeded with our experiments by applying first the lexical rules, and subsequently the contextual rules. At each point (T_{start} and T_{final}) we measured the tagging accuracy, in order to assess the respective contributions from the lexical and contextual rules. This procedure is illustrated in Figure 7. Experimental results on the test sets are shown in Figure 8.

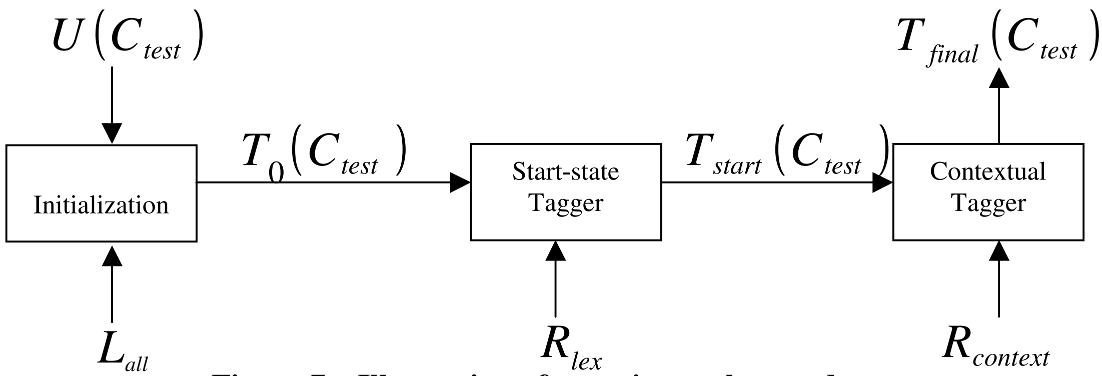

Figure 7. Illustration of experimental procedure.

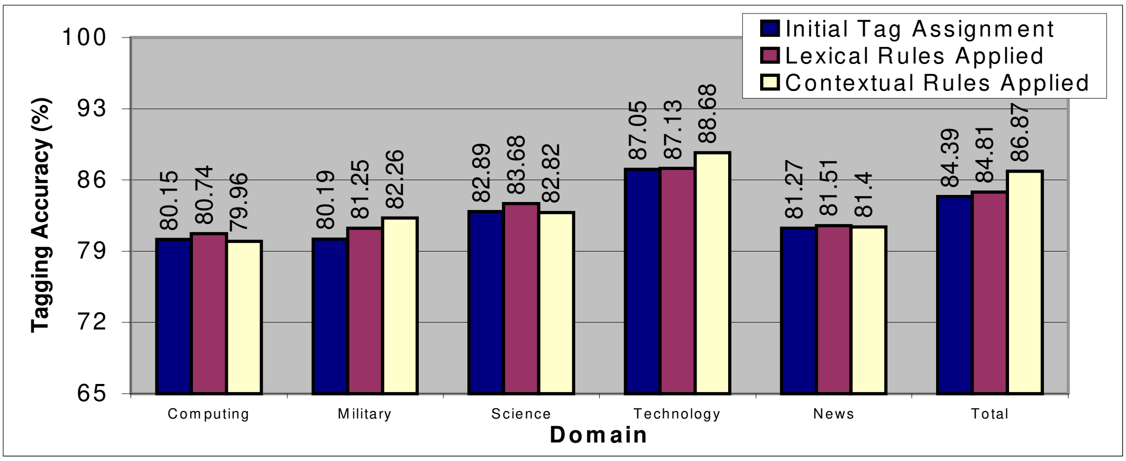

Figure 8. Tagging accuracies on the test sets.

Figure 8 shows that the lexical rules brought about a small but consistent improvement (from 0.08% to 1.06% across different domains) over the initial tag assignments across all the domains. However, the contextual rules led to a slight degradation in performance in three of the five domains. For the "Total" category, we believe that the relatively higher improvement is due to a greater amount of training data made available from gathering together 90% of the entire corpus and the co-operation between lexical rules and contextual rules. As an illustration of the co-operation between lexical rules and contextual rules, consider the example sentence:

Untagged Sentence: 各個崗位上的各族青年朋友致以節日的祝賀 / !

Reference Sentence: 各/rn 個/qng 崗位/ng 上/f 的/usde 各/rn 族/ng 青年/ng 朋友/ng 致以/vgn

節日/ng 的/usde 祝賀/vg ! / !

Since 致以 is an unknown word, which is tagged as ng by the start-state tagger. After the initial tag assignments and application of the lexical rule {以 hassuf 2 vgv}, the sentence is tagged as:

各/rn 個/qng 崗位/ng 上/f 的/usde 各/rn 族/ng 青年/ng 朋友/ng 致以/vgv 節日/ng 的/usde 祝賀/vg ! / !

Finally, the application of the contextual rule {vgv vgn SURROUNDTAG ng ng} corrects the tag for 致以 from vgv to vgn and it's the correct tag for 致以 in the sentence.

In order to further assess the contribution of the contextual rules, we examined their effects on the training corpus. Results are shown in Figure 9. Since the training corpus does not have unknown words, we only have two sets of tagging accuracies - one from the initial tag assignments, and the other from lexical rule application.

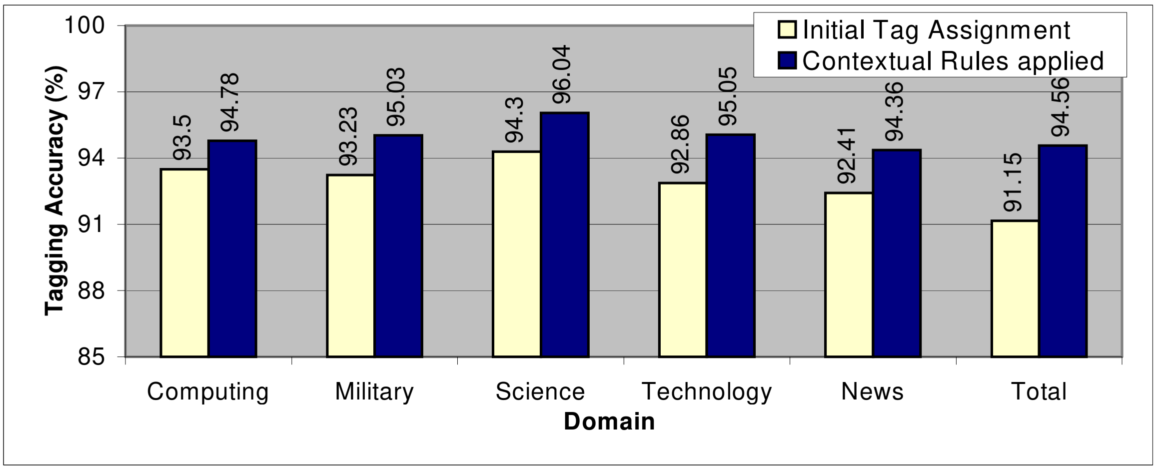

Figure 9. Tagging accuracies (%) on the training sets.

For the results in Figure 9, the initial tag assignments utilized the lexicon derived from the training set of the corresponding domain only. Compared to the test-set results, the contextual rules contributed to a more pronounced improvement, across the training sets in all the domains. The improvement did not carry over to the test sets, possibly due to over-fitting to the training sets.

4.3 Performance on Unknown Words

We have also examined our tagging performance on the unknown words and unknown tags in the test set. Performance accuracies on unknown words range between 40 to 50%, as shown in Table 5.

Test Sets	Computing	Military	Science	Technology	News	Total
Unknown word Performance	55.41	44.73	56.16	53.33	43.31	56.57

Table 5. Tagging accuracies (%) on the unknown words in the test sets.

Our experiments have also shown that the contextual rules learnt have not corrected any of the unknown tag errors in the test set. One reason is due to the propagation of errors - an errorful tag assignment to an unknown word may propagate via contextual rule applications to cause errors in subsequent tags. As an illustration of error propagation, consider the example sentence:

全市鄉鎮企業中已有30 多家中外合資合作企業。

where 合資 is the unknown word, the tag **qni** (個體量詞) is the unknown tag. After the initial tag assignments and application of the lexical rules, the sentence is tagged as:

全/a 市/ng 鄉鎮/ng 企業/ng 中/f 已/d 有/vh 30/mx 多/mg 家/ng 中/f 外/f 合資/ng 合作/vg 企業/ng 。/。

The unknown word 合資 tagged as **ng**.

Subsequent to this, application of the contextual rule {vg vgn prev1or2tag ng} transforms the tag for 合作 (from **vg** to **vgn**) since its left tag of word 合資 **ng**. Therefore, the tag of 合作 becomes an error. Now the sentence tags become:

全/a 市/ng 鄉鎮/ng 企業/ng 中/f 已/d 有/vh 30/mx 多/mg 家/ng 中/f 外/f 合資/ng 合作/vg 企業/ng 。/。

This is compared with the reference tags:

全/a 市/ng 鄉鎮/ng 企業/ng 中/f 已/d 有/vh 30/mx 多/mg 家/ng 中/f 外/f 合資/ng 合作/vg 企業/ng 。/。

We find five errors in the TEL tagging:

家/ng, 中/f, 外/f, 合資/ng, 合作/ng (hypothesized)

家/qni, 中/j, 外/j, 合資 合作/g (reference)

and among these three originated from unknown words and unknown tags (家, 合資, 合作

4.4 A Possible Benchmark

We attempt to come up with an upper bound benchmark for our performance accuracies, by ameliorating the unknown word problem. To achieve this we included all the words in our *entire* corpus (L_{all}) for initial tag assignment. We have also used the entire training corpus for training the contextual rules (instead of divided it into the lexical and contextual portions, as mentioned previously). This experimental procedure is illustrated in Figure 10. Our experimental results suggest that possible upper bounds for tagging performance lies around 97% for training and 94% for testing in domain total. This compares with the previous performances of 94.56% in the training set (please see Figure 9 in pp.16) and 86.87% in the testing set (please see Figure 8 in pp. 16).

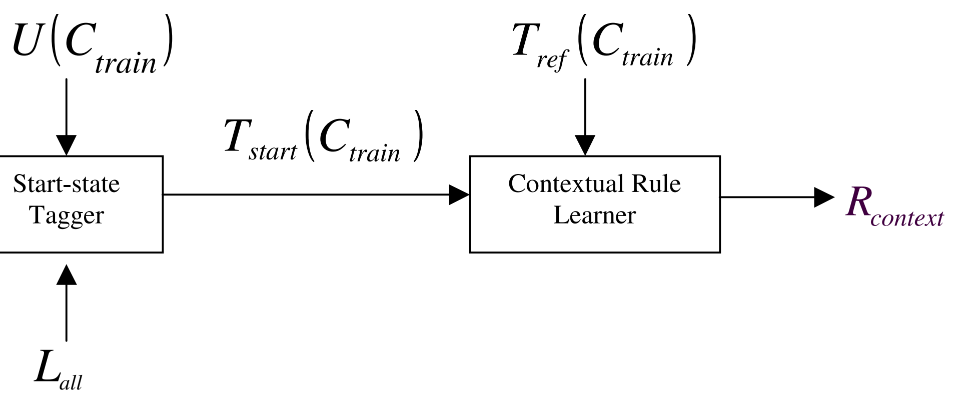

Figure 10. Training procedure which attempts to ameliorate the effect of unknown words.

Experimental results for both training and test sets are tabulated in Table 6.

Domains	Computing	Military	Science	Technology	News	Total
Training Accuracies	96.13	96.98	97.70	96.98	96.70	96.96

Testing Accuracies	94.10	92.05	94.18	92.51	92.73	93.88

Table 6. Tagging accuracies (%) for both training and test sets, under the condition with no unknown words.

4.5 Comparison between the TEL approach and the stochastic approach

We attempted to compare the TEL approach with a stochastic approach for POS tagging. Our stochastic tagger is provided by Tsinghua University. It utilizes a Markov model for POS tagging, i.e.

$$P(T's \mid W_s) = \max_{T_1 T_2 \ldots T_n} P(T_1 \mid T_2) \prod_{i=2}^{n} P(T_i \mid T_{i-1}) P(W_i \mid T_i)$$

and has been previously trained.[6] Therefore it was not straightforward for us to compare the two taggers based on identical training and testing sets. We divided each corpus into 10 partitions – 9 of them were used to train the TEL tagger and the remaining one for testing. This preserves the 9:1 divide between training and testing sets. These experiments are repeated 5 times by jackknifing the data sets, and the performance accuracies were averaged (see row 2, row 3 and column 7 of Table 7). We combined the average training and testing accuracies according to the formula:

Overall Accuracy (TEL) = 0.9 x average training accuracy + 0.1 average testing accuracy

The weights of the training and testing accuracies follow the proportion of the respective data sets. The Overall Accuracy (TEL), shown in the third row of Table 7 were compared with the corresponding values of the stochastic tagger, shown in the last row of the table. Our results suggest that the TEL and stochastic approaches produce comparable results.

Experimental Runs	1	2	3	4	5	Average (over 5 runs)

[6] Previous literature indicates that the training was based on 90% of the corpus.

TEL tagger (Training Accuracy)	95.20	95.17	95.16	95.00	95.17	95.14
TEL tagger (Testing Accuracy)	88.33	87.60	87.46	88.40	87.26	87.80
TEL tagger (Overall Accuracy)	94.50	94.35	94.33	94.39	94.41	94.38
Tsinghua tagger	91.59	91.59	91.59	91.59	91.59	91.59

Table 7. Tagging accuracies (%) for both training and test sets. Comparison between the TEL approach and stochastic approach.

5. Conclusion

This work is our initial attempt in using the transformation-based error-driven learning (TEL) procedure for tagging Chinese text. TEL has previously been shown to be effective in POS tagging for English (achieving over 96% tagging accuracies in using the Brown and WSJ corpora) [Brill 1995]. It has several attractive properties: (i) it provides an automatic procedure for tagging, (ii) the lexical and contextual rules it learns often make intuitive sense for the Chinese language, and potential provides room for the incorporation of linguistic knowledge by a human, should there be sparse training data problems, (iii) the learning procedure aims to minimize errors to obtain maximum tagging accuracies.

Using a Chinese news corpus of over 70,000 words, divided into disjoint training and test sets of a 9:1 ratio, we achieved overall tagging accuracies of 94.56% (training) and 86.87% (testing). Across the different domains, the proportion of unknown words and unknown tags range between 8% to 33%, and tagging performance from 79.96% to 88.68%. In general, the higher the proportion of unknown words/tags, the lower the tagging performance. The baseline performance (without applying any rules) was 91.16% (training) and 84.39% (testing). Both the lexical and contextual rules were found to be contributive towards tagging performance. Performance accuracies are much improved upon the use of a comprehensive lexicon to ameliorate the unknown word problem, reaching 96.96% (training) and 93.88% (testing) respectively as a possible gauge of an upper bound performance for our experiment. While direct comparison with the work of others[7] is difficult due to uncertainties in training/testing data partitioning, our experimental results in comparison with a stochastic tagger suggests that TEL is equally effective and applicable for Chinese.

Acknowledgements

We thank Eric Brill for his transformational tagger, and Tsinghua University (in Beijing) for providing the corpora for our experiments.

References

Bai, S. H., Xia, Y. and Huang, C. N., "Automatic Part of Speech Tagging System for Chinese" , Technical Report, Tsinghua University, Beijing, China, 1992.

Black, E., A. Finch and H. Kashioka, "Trigger-Pair Predictors in Parsing and Tagging", Proceedings of the International Conference on Computational Linguistics, pp. 131-137, 1998.

Brill, E., "Transformation-based Error-driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging", Computational Linguistics, Vol. 21, Number 4, 1995.

Chang, C. H. and Chen, C. D., "A Study on Integrating Chinese Word Segmentation on Part-of-Speech Tagging", Communications of COLIPS, Vol. 3, No. 1, pp. 69-77, 1993.

Chiang, T. H., Chang, J. S., Lin M. Y. and Su K. Y., "Statistical Word Segmentation", Journal of Chinese Linguistics, pp. 147-174, 1996.

Chen, C.J., M.H. Bai, K.J. Chen, 1997, "Category Guessing for Chinese Unknown Words" Proceedings of the Natural Language Processing Pacific Rim Symposium 1997: pp. 35-40, NLPRS1997 Thailand.

Church, K., "A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text", Proceedings of ICASSP-89, pp. 695-698, 1989.

Jelinek, F., "Self-Organized Language Modeling for Speech Recognition", Readings in Speech Recognition, A. Waibel and K. F. Lee, eds., Morgan Kaufman Publishers, 1990.

Kupiec, J., "Robust Part-of-Speech Tagging using a Hidden Markov Model", Computer Speech and Language, 6:226-242, 1992.

Lua, K. T., "Part of Speech Tagging of Chinese Sentences using Genetic Algorithm", International Conference on Chinese Computing, pp. 45-49, 1996.

[7] e.g. [Bai et al., 1992]

Merialdo, B., "Tagging Text with a Probabilistic Model", Proceedings of ICASSP-91, pp. 809-812, 1991.

Qin A. and Wong, W. S., "ACCESS: Automatic Segmentation and Part of Speech Tagging of Chinese Text", Technical report, The Chinese University of Hong Kong, 1998.

Shing-Huan Liu, Keh-jiann Chen, Li-ping Chang, Yeh-Hao Chin, "A Practical Tagger for Chinese Corpora", Proceedings of ROCLING VII, pp.111-126

Su, K. Y., Chiang, T. H. and Chang, J.S., "An Overview of Corpus-based Statistics-Oriented (CBSO) Techniques for Natural Language Processing", Computational Liguistics and Chinese Language Processing, Vol. 1., No. 1, pp. 101-157, August 1996.