REPAIRING ERRORS FOR CHINESE WORD SEGMENTATION AND PART-OF-SPEECH TAGGING

TIAN-FANG YAO , WEI DING , GREGOR ERBACH

Computational Linguistics Department, Saarland University, Germany.
E-mails: yao@coli.uni-sb.de , wding@dfki.de , gor@acm.org

Abstract:
For improving the recognition performance of Chinese named entities, transformation based machine learning has been introduced to repair errors caused during word segmentation and part-of-speech (POS) tagging. Because Chinese is not a segmented language, the words in a sentence must be segmented before they are processed by consequent Chinese named entity recognition component. Similarly, POS tagging is also an important fundamental task for Chinese named entity recognition. In order to enhance the quality of word segmentation and POS tagging, it is necessary to explore different approaches for improving the performance. One of the approaches is to repair errors as much as possible, if a word segmentation and POS tagging tool is available on hand. This paper aims to introduce an effective error repairer using transformation based error-driven machine learning technique. It deals with detecting error positions, producing error repairing rules, selecting higher-score rules, ordering rules and distinguishing rule usage conditions etc. The experimental results show that word segmentation and POS tagging errors are significantly reduced and the performance has been improved. The average recall for word segmentation and POS tagging has gone up by 4.01% and 11.98% respectively; while the average precision for word segmentation and POS tagging has gone up by 6.07% and 13.02% respectively. At the same time, another experimental results have also shown that the average F-measure for the recognition of six Chinese named entities has also increased by 14.04% [8].

Keywords:
Transformation based machine learning, Transformation based error-driven machine learning, repairing errors for Chinese word segmentation and part-of-speech tagging, Chinese named entity recognition

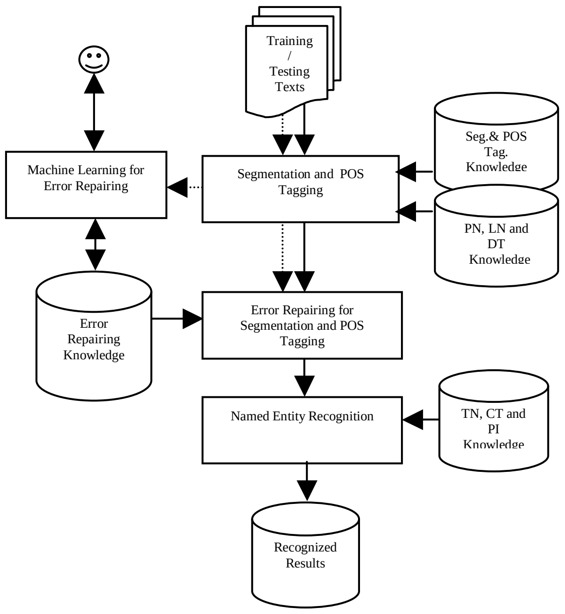
1 Introduction

Chinese word segmentation and POS tagging is a basic task for Chinese language processing. This component is a “bottle neck”, the quality of which has a direct impact on the performance of consequent processing components. The quality measure relates to the performance evaluation (recall, precision and F-measure) of word segmentation and POS tagging. Therefore, if the quality of word segmentation and POS tagging system is not adequate, it can be beneficial to develop an effective stand-alone error repairer to reduce errors as much as possible.

In order to guarantee good quality of word segmentation and POS tagging, we compared different existing word segmentation and POS tagging systems and introduced Modern Chinese Word Segmentation and POS Tagging System [1] as the first processing component in our Chinese named entity recognition model. In [1], the word segmentation part is based on AB (Association-Backtracking) algorithm and mainly depends on Chinese language knowledge, such as word-building, form-building and syntax. It adopts practical word segmentation rules to resolve ambiguous structure problems. In addition, the POS tagging part utilizes the probability statistic model as well as CLAWS, VOLSUNGA and the corresponding transmutation algorithms.

Unfortunately, we found there exist numerous word segmentation and POS tagging errors when we make use of this system to process our texts from the sports domain. Apparently, these errors will have an effect on the consequent recognition for the Chinese named entities. The main work of this paper is to investigate an error repairer for reducing errors from word segmentation and POS tagging, so that it can improve the overall performance of word segmentation and POS tagging system. Our work includes designing the algorithms for automatic error repairing, defining rule formats, distinguishing context sensitive or context free rules against different error situations, and showing the experimental results to compare the performance between word segmentation and POS tagging system and the consequent error repairer.

Figure 1 is the model of our Chinese named entity recognition. The dotted line shows the flow process for the training texts; while the solid line is one for the testing texts. When training, the texts are segmented and tagged, then the error repairing rules are produced and some of them are selected as the regular rules under the appropriate conditions. Thereafter, the errors caused during word segmentation and POS tagging in testing texts can be automatically corrected through utilizing such regular rules. Among the six named entities, personal name (PN), date and time (DT) as well as location name (LN) are tagged by the word segmentation and POS tagging component and corrected by the error repairer; while team name (TN), competition title (CT) and personal identity (PI) will be tagged by the named entity recognition component.



Training / Testing Texts
Seg.& POS Tag. Knowledge
Machine Learning for Error Repairing
Segmentation and POS Tagging
PN, LN and DT Knowledge
Error Repairing Knowledge
Error Repairing for Segmentation and POS Tagging
Named Entity Recognition
TN, CT and PI Knowledge
Recognized Results

Fig 1. Chinese named entity recognition model

This paper is organized as follows. Section 2 illustrates the error types and the definition of error repairing rules, section 3 presents machine learning algorithm for training and testing, section 4 shows the experimental results, and the final section draws some conclusions.

2 Error type and error repairing rule

In the output of the system [1], the word segmentation and POS tagging errors have different types. Some examples are shown as follows:

Word segmentation errors

- An abbreviated word is combined with other words,
- An enclitic is segmented with a numeral together,
- An idiom is segmented into some parts,
- A general noun is segmented into some parts or combined with other words,
- A location name, a personal name or a team name is segmented into some parts or
- A verb is segmented with an auxiliary word or an adverb together, etc.

POS tagging errors

- An abbreviated word is tagged as a general noun,
- An adverb is tagged as an adjective or a quantifier,
- An auxiliary word is tagged as a verb,
- A general noun is tagged as a personal name, a verb or a location name,
- A numeral is tagged as a general noun or an adverb,
- A person name is tagged as a location name,
- A pronoun is tagged as a general noun or
- A verb is tagged as a general noun, etc.

Examples:

Ex1. 上视|J|又|D|利|N|滋|G|女足|J
“又利滋” (You Li Zi) is a team name which should not be segmented.

Ex2. 姑娘|N|狂|A|轰|V|乱|A|炸|V|，|W
“狂轰乱炸” (bomb savagely) is an idiom which should also not be segmented.

Ex3. 球|N|小胜|N5|大连|N5
“小胜” (Xiao Sheng) is not a Chinese location name, but it should be segmented into two words. The word “小” (by the close score) is an adverb; while the word “胜” (win) is a verb.

Ex4. 以一记|N7|有力|A
“以一记” (Yi Yi Ji) is not a transliterated personal name or location name, but it should be segmented into three words. The first word “以” (with) is a proposition, the second word “一” (a) is a numeral and the third word “记” (slap) is a quantifier.

Ex5. 特别|D|提|V|到了|D|年轻|A
“提到了” (have mentioned) should not be segmented into “提” (carry; lift; raise etc.) and “到了” (in the end; finally; eventually etc.), but it should be segmented into “提到” (mention) and “了” (auxiliary word). The word “提到” is a verb, the word “了” is an auxiliary word.

Ex6. 他|R|在场|V|上前|V|冲|N|后|D|堵|V|的|U
“在场上前冲后堵” (attack in front and guard behind on the ground) should be segmented into “在” (on), “场上” (ground) and “前冲后堵” (attack in front and guard behind). The first word is a proposition, the second word is a general noun and the third word is an idiom.

Ex7. 本|N|报|N
The word “本” (one’s own; native) is not a general noun, but it is an adjective.

Note: In the above examples, A, D, G, J, N, N5, N7, R, U, V and W represent an adjective, an adverb, a morpheme, an abbreviation, a general noun, a LN, a transliterated PN or LN, a pronoun, an auxiliary word, a verb and a punctuation respectively.

Following [2, 3] we divide the error repairing operations of word segmentation into three types, that is, **concat**, **split** and **slide**. **concat** means that some characters (or words) that have been separated will be put together; **split** means that some characters (or words) that have been put together will be separated and **slide** denotes that some characters (or words) whose segmentation positions are not correct will be segmented newly. That is, the new position will be moved to the left or right side of the original position. The three types of error repairing rule for word segmentation are defined as follows:

rectify_segmentation_error (**concat**, old_word1 | old_tag1 | old_word2 | old_tag2 | …, concat_number, new_tag, preceding_word | preceding_tag, following_word | following_tag)

rectify_segmentation_error (**split**, old_word | old_tag, split_position1 | split_position2 | …, new_tag1 | new_tag2 | …, preceding_word | preceding_tag, following_word | following_tag)

rectify_segmentation_error (**slide**, old_word1 | old_tag1 | old_word2 | old_tag2 | …, slide_direction_length1 | slide_direction_length2 | …, new_tag1 | new_tag2 | …, preceding_word | preceding_tag, following_word | following_tag)

In the above error repairing rules, the preceding word and tag as well as the following word and tag mean the context of the error segmentation for word or phrase. If the concrete rule is a context-free rule, this part is omitted. The old word or tag denotes the word or tag before repairing errors; while the new word or tag represents the word or tag after repairing errors. The **concat** number is the number of **concat** operations. The **split** position is the position between characters from left to right in the ascending order. The **slide** direction is either left or right. The **slide** length is calculated by character number.

In addition, we also define the error repairing rule for POS tagging:

rectify_tag_error (old_word | old_tag, new_tag, preceding_word | preceding_tag, following_word | following_tag)

The following are the examples of error repairing rule. They correspond to above examples:

Ex1. rectify_segmentation_error (**concat**, 又|D|利|N|滋|G, 2, N, 上视|J, 女足|J)

Ex2. rectify_segmentation_error (**concat**, 狂|A|轰|V|乱|A|炸|V, 3, I, 姑娘|N, ，,|W)

Ex3. rectify_segmentation_error (**split**, 小胜|N5, 1, D|V, 球|N, 大连|N5)

Ex4. rectify_segmentation_error (**split**, 以一记|N7, 1|2, P|M|Q, _|_,有力|A)

Ex5. rectify_segmentation_error (**slide**, 提|V|到了|D, right1, V|U, 特别|D, 年轻|A)

Ex6. rectify_segmentation_error (**slide**, 在场|V|上前|V|冲|N|后|D|堵|V, left1|left1|right2|right1, P|N|I, 他|R, 的|U)

Ex7. rectify_tag_error (本|N, A, _|_,报|N)

In the above rules, we define convenient error repairing operations which can be suitable to process different error types. In addition, we add context constraints in the rule for distinguishing different error situations in the use.

3 Error repairing algorithm

For word segmentation and POS tagging errors, we use machine learning technique to teach the system how it can recognize and repair such errors. We mainly adopt transformation-based error-driven machine learning method [2, 3, 4] that is based on a text corpus. It has been applied to different applications, e.g. part-of-speech tagging [5], prepositional phrase attachment disambiguation [6], bracketing text [7], etc.

We deal with different error cases in different context constraints of error repairing rules in our algorithm during testing period. That is, the rules are divided into three types manually: whole POS context constraints, preceding POS context constraints and without context constraints. In the former two types of rules, the rules don’t be suitable to all contexts. Therefore, we must limit their usage through POS of the context.

For example, in the sentence "上半场第 8 分钟中远队就发起了进攻。" (At the 8th minute in the first half-time, Zhongyuan Team launched an attack immediately), "中远" (Zhongyuan) is a team name which should not be separated. We can use the **concat** rule "rectify_segmentation_error (**concat**, 中|F|远|A, 1, N, 分钟|Q, 队|N)" to put two characters together. In the rule, F and Q represent a direction word and a quantifier separately. In the another sentence "他在带球突破中远射破门。" (He shot a goal at a long distance and broke the goal during he dribbled the ball and broke through.) "中远" is two separated characters "中" (during) and "远" (at a long distance). They should not be put together. Because we use POS context constraints in the above rule, the POS tag of "突破" (break through) and "射" (shoot) is V. "中远" in second sentence will be separated. After these two sentences have been repaired, the output from the error repairer is respectively displayed as follows:

上半场|T|第|H|8|M|分钟|T|中远|N|队|N|就|D|发起|V|了|U|进攻|V|。|W|

他|R|在|P|带球|V|突破|V|中|N|远射|V|破门|V|。|W|

Note that H, M, P and T mean a prefix, a numeral, a proposition and date or time respectively. The POS tag of "分钟" (minute) is T because the tag has been repaired.

The following are the concrete training and testing algorithm:

Training Algorithm

1) Compare automatic word segmentation and POS tagging with manual word segmentation and POS tagging in a sentence.
2) If they are different, record word segmentation and POS tagging environments. Otherwise transfer to 6).
3) Build a new transformation rule that consists of transformation condition and action. The condition represents all triggering environment including error word segmentation position, error POS and context. The action executes repairing action that transforms word segmentation positions and POS tags.
4) Examine whether this new transformation rule is at odds with the transformation rules in the candidate rule library. If it is true, either merge rules or delete this new rule depending on both conditions. Otherwise add the new rule into the library.
5) Test the rules in the candidate rule library and choose some higher-score transformation rules. The score is the repaired error number for a rule in the training set. Then determine whether they are added into the final rule library.
6) If there is still a sentence to be processed in a text, transfer to 1).
7) Order the rules depending on their score.

In the training algorithm, the error positions are determined by comparing between manually processed text and automatically processed text. Simultaneously, the error environments are recorded. Based on such information, the transformation rules are generated and the effective error repairing rules are selected. These rules can repair more than one error in the training set. In order to use these rules with priority, the rule in the final rule library are sorted.

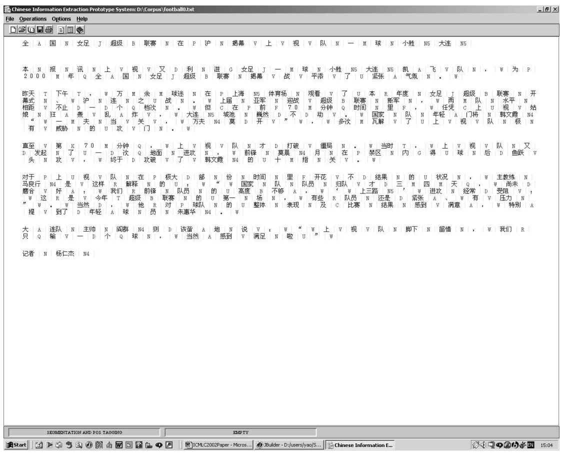
Testing Algorithm

1) Input a sentence by automatic word segmentation and POS tagging.
2) Repair word segmentation errors. In the repairing, retrieve the final rule library in such sequence: whole POS context constraints, preceding POS context constraints and without context constraints. If one of rules in rule library is matched, execute the corresponding transformation action.
3) Repair POS tagging errors. Its error repairing procedure is same as 2).
4) If there is still a sentence to be processed, transfer to 1).

In the testing algorithm, the usage of error repairing rules with context constraints are prior to those without context constraints, the employment of error repairing rules for word segmentation have priority over those for POS tagging. Thus, it ensures that the rules repair much more errors. At the same time, it prevents new errors occur during repairing existing errors.

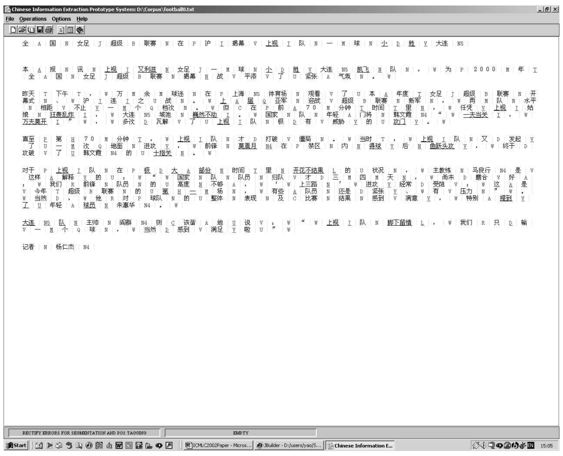
4 Experimental results

Chinese Named Entity Recognition System (CNERS) has been implemented with Java 2 (ver.1.4.0) under Windows 2000. The error repairer is integrated as a component in the system. Figure 2, 3 and 4 show the user interfaces of the System CNERS. The user interfaces display the result for word segmentation and POS tagging, the result for error repairing as well as the statistic report for error repairing respectively. In Figure 2, the black characters represent Chinese characters or words, and the red alphabetical characters denote POS tags. In Figure 3, the black underlines and the red underlines represent the repaired part for Chinese character or word and POS tag separately. In Figure 4, it shows the error repairing number for word segmentation and POS tagging.



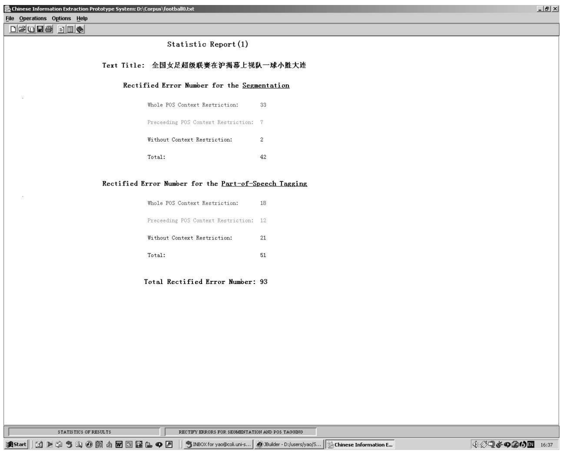
```
Chinese Information Extraction Prototype System: D:\Corpus\football0.txt
File  Operations  Options  Help

全 A 国 N 女足 J 超级 B 联赛 N 在 P 沪 N 揭幕 V 上 V 视 V 队 N 一 M 球 N 小胜 N5 大连 N5

本 N 报 N 讯 N 上 V 视 V 又 D 利 N 滋 G 女足 J 一 M 球 N 小胜 N5 大连 N5 凯 A 飞 V 队 N ， W 为 P
2000 M 年 Q 全 A 国 N 女足 J 超级 B 联赛 N 揭幕 V 战 V 平添 V 了 U 紧张 A 气氛 N 。 W

昨天 T 下午 T ， W 万 M 余 M 球迷 N 在 P 上海 N5 体育场 N 观看 V 了 U 本 R 年度 N 女足 J 超级 B 联赛 N 开
幕式 N 、 W 沪 N 连 N 之 U 战 N 。 W 上届 N 亚军 N 迎战 V 超级 B 联赛 N 新军 N ， W 两 M 队 N 水平 N
相距 V 不止 D 一 D 个 Q 档次 N 。 W 但 C 在 P 前 F 70 M 分钟 Q 时间 N 里 F ， W 任凭 C 上 U 视 V 姑
娘 N 狂 A 轰 V 乱 A 炸 V ， W 大连 N5 城池 N 巍然 D 不 D 动 V 。 W 国家 N 队 N 年轻 A 门将 N 韩文霞 N4
“ W 一 M 夫 N 当 V 关 V ， W 万夫 N4 莫 D 开 V ” W ， W 多次 M 瓦解 V 了 U 上 V 视 V 队 N 极 N
有 V 威胁 N 的 U 攻 V 门 N 。 W

直至 V 第 K 70 M 分钟 Q ， W 上 V 视 V 队 N 才 D 打破 V 僵局 N 。 W 当时 T ， W 上 V 视 V 队 N 又
D 发起 N 了 U 一 D 次 Q 地面 N 进攻 N ， W 前锋 N 莫晨 N4 月 N 在 P 禁区 N 内 G 得 U 球 N 后 D 鱼跃 V
头 N 攻 V ， W 终于 D 攻破 V 了 V 韩文霞 N4 的 U 十 M 指 N 关 V 。 W

对于 P 上 U 视 V 队 N 在 P 极大 D 部 N 份 N 时间 N 里 F 开花 V 不 D 结果 N 的 U 状况 N ， W 主教练 N
马良行 N4 是 V 这样 R 解释 N 的 U ： W “ W 国家 N 队 N 队员 N 归队 V 才 D 三 M 四 M 天 Q ， W 尚未 D
磨合 V 好 A ； W 我们 R 前锋 N 队员 N 的 U 高度 B 不够 A ， W ‘ W 上三路 N5 ’ W 进攻 N 经常 D 受阻 V ；
W 这 R 是 V 今年 T 超级 B 联赛 N 的 U 第一 N 场 N ， W 有些 R 队员 N 还是 D 紧张 A 、 W 有 V 压力 N
” W 。 W 当然 D ， W 他 R 对 P 球队 N 的 U 整体 N 表现 N 及 C 比赛 N 结果 N 感到 V 满意 A ， W 特别 A
提 V 到了 D 年轻 A 球 N 员 N 朱惠华 N4 。 W

大 A 连队 N 主帅 N 阎群 N4 则 D 诙谐 A 地 N 说 V ： W “ W 上 V 视 V 队 N 脚下 N 留情 N ， W 我们 R
只 Q 输 V 一 D 个 Q 球 N ， W 当然 A 感到 V 满足 N 啦 U ” W

记者 N 杨仁杰 N4

SEGMENTATION AND POS TAGGING | EMPTY
```

Fig 2. User interface for word segmentation and POS tagging



```
Chinese Information Extraction Prototype System: D:\Corpus\football0.txt
File  Operations  Options  Help

全 A 国 N 女足 J 超级 B 联赛 N 在 P 沪 I 揭幕 V 上视 I 队 N 一 M 球 N 小 D 胜 V 大连 N5

本 A 报 N 讯 N 上视 I 又利滋 N 女足 J 一 M 球 N 小 D 胜 V 大连 N5 凯飞 N 队 N ， W 为 P 2000 M 年 T
全 A 国 N 女足 J 超级 B 联赛 N 揭幕 N 战 V 平添 V 了 U 紧张 A 气氛 N 。 W

昨天 T 下午 T ， W 万 M 余 M 球迷 N 在 P 上海 N5 体育场 N 观看 V 了 U 本 A 年度 T 女足 J 超级 B 联赛 N 开
幕式 N 、 W 沪 I 连 I 之 U 战 N 。 W 上 A 届 Q 亚军 N 迎战 V 超级 B 联赛 N 新军 N ， W 两 M 队 N 水平
N 相距 V 不止 V 一 M 个 Q 档次 N 。 W 但 C 在 P 前 A 70 M 分钟 T 时间 T 里 N ， W 任凭 V 上视 I 姑
娘 N 狂轰乱炸 I ， W 大连 N5 城池 N 巍然不动 I 。 W 国家 N 队 N 年轻 A 门将 N 韩文霞 N4 “ W 一夫当关 I ， W
万夫莫开 I ” W ， W 多次 D 瓦解 V 了 U 上视 I 队 N 极 D 有 V 威胁 V 的 U 攻门 V 。 W

直至 P 第 H 70 M 分钟 T ， W 上视 I 队 N 才 D 打破 V 僵局 N 。 W 当时 T ， W 上视 I 队 N 又 D 发起 V
了 U 一 M 次 Q 地面 N 进攻 V ， W 前锋 N 莫晨月 N4 在 P 禁区 N 内 N 得球 V 后 N 鱼跃头攻 V ， W 终于 D
攻破 V 了 U 韩文霞 N4 的 U 十指关 N 。 W

对于 P 上视 I 队 N 在 P 极 D 大 A 部份 N 时间 T 里 N 开花不结果 L 的 U 状况 N ， W 主教练 N 马良行 N4 是 V
这样 A 解释 V 的 U ： W “ W 国家 N 队 N 队员 N 归队 V 才 D 三 M 四 M 天 N ， W 尚未 D 磨合 V 好 A
； W 我们 R 前锋 N 队员 N 的 U 高度 N 不够 A ， W ‘ W 上三路 N ’ W 进攻 V 经常 D 受阻 V ； W 这 A 是
V 今年 T 超级 B 联赛 N 的 U 第 H 二 M 场 N ， W 有些 A 队员 N 还是 D 紧张 V 、 W 有 V 压力 N ” W 。
W 当然 D ， W 他 R 对 P 球队 N 的 U 整体 N 表现 N 及 C 比赛 N 结果 N 感到 V 满意 V ， W 特别 A 提到 V
了 U 年轻 A 球员 N 朱惠华 N4 。 W

大连 N5 队 N 主帅 N 阎群 N4 则 C 诙谐 A 地 U 说 V ： W “ W 上视 I 队 N 脚下留情 L ， W 我们 R 只 D 输
V 一 M 个 Q 球 N ， W 当然 D 感到 V 满足 V 啦 U ” W

记者 N 杨仁杰 N4

RECTIFY ERRORS FOR SEGMENTATION AND POS TAGGING | EMPTY
```

Fig 3. User interface for error repairing



```
Chinese Information Extraction Prototype System: D:\Corpus\football0.txt
File  Operations  Options  Help

                    Statistic Report(1)

        Text Title:  全国女足超级联赛在沪揭幕上视队一球小胜大连

          Rectified Error Number for the Segmentation

               Whole POS Context Restriction:       33

               Preceeding POS Context Restriction:  7

               Without Context Restriction:         2

               Total:                               42

        Rectified Error Number for the Part-of-Speech Tagging

               Whole POS Context Restriction:       18

               Preceeding POS Context Restriction:  12

               Without Context Restriction:         21

               Total:                               51

             Total Rectified Error Number: 93

STATISTICS OF RESULTS | RECTIFY ERRORS FOR SEGMENTATION AND POS TAGGING
```

Fig 4. User interface for statistic report of error repairing

The training set consists of 94 texts including 3473 sentences (roughly 37077 characters) from Jie Fang Daily (http://www.jfdaily.com/) in 2001. The texts come from football sports news. After learning, we obtain 4304 transformation rules. Among them 2491 rules are for word segmentation error repairing, 1813 rules are for POS tagging error repairing. There are 1730 rules as **concat** rules, 554 rules as **split** rules and 207 rules as **slide** rules in the word segmentation error repairing rules. Subsequently, we distinguish above rules into context-sensitive or context-free category manually. In the word segmentation rules, 790, 315 and 77 rules are as **concat, split** and **slide** context-sensitive rules respectively; while 940, 239 and 130 rules are as **concat, split** and **slide** context-free rules separately. In the POS tagging rules, 1052 rules are context-sensitive rules and 761 are context-free rules.

The testing set is a separate set, which contains 20 texts including 658 sentences (roughly 8340 characters). The texts in the testing set have been randomly chosen from Jie Fang Daily in May 2002. The texts also come from football sports news.

The evaluation for the performance of the error repairer is composed of four measures, that is, recall, precision, F-measure and error repairing rate. Based on [2], we define recall, precision, F-measure and error repairing rate as follows:

$$\text{Recall} = \frac{\text{correct machine-segmented (tagged) word number}}{\text{hand-segmented (tagged) word number}}$$

$$\text{Precision} = \frac{\text{correct machine-segmented (tagged) word number}}{\text{machine-segmented (tagged) character and word number}}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{Error Repairing Rate} = \frac{\text{repaired machine-segmented (tagged) error number}}{\text{total machine-segmented (tagged) error number}}$$

Table 1, 2 and 3 show the experimental results for the performance in different cases.

These results indicate that the average F-measure of word segmentation has increased by 5.11%; while one of POS tagging has even increased by 12.54%. We have shown that the learning and error repairing algorithm for POS tagging is more effective.

	Ave. Recall	Ave. Precision	Ave. F-measure
Without Error Repair	91.07	84.67	87.75
With Error Repair	95.08	90.74	92.86

Tab 1. Performance for word segmentation

	Ave. Recall	Ave. Precision	Ave. F-measure
Without Error Repair	80.41	74.73	77.47
With Error Repair	92.39	87.75	90.01

Tab 2. Performance for POS tagging

	Ave. Error Repairing Rate
Segmentation	39.99
POS Tagging	56.68

Tab 3. Average error repairing rate for word segmentation and POS tagging

The paper [8] has also given another experimental results related with this work. Using same testing set, the performance for most of named entities in our system has manifestly been improved, the average F-measure for the recognition of six named entities has increased by 14.04% from 61.30% to 75.34%.

5 Conclusions

In Chinese information extraction investigation, we noted that the errors from word segmentation and POS tagging have adversely affected the performance of named entity recognition to a certain extent. We utilize transformation based error-driven machine learning to perform error repairing for word segmentation and POS tagging simultaneously. In error repairing procedure, we add context-sensitive or context-free constraints in the rules. Thus it can avoid the introduction of further errors during error repairing. The experimental results have shown that the performance of word segmentation and POS tagging has been improved, leading to an improved recognition performance for named entities in our system. Such a hybrid approach used in our system synthesizes the advantages of knowledge engineering and machine learning.

Acknowledgements

This work is a part of the COLLATE (Computational Linguistics and Language Technology for Real World Applications) project under contract no. 01INA01B, which is being supported by the German Ministry for Education and Research.

References

[1] K.Y. Liu, Automatic Segmentation and Tagging for Chinese Text, The Commercial Press, Beijing, China, (In Chinese), 2000

[2] D. Palmer, A Trainable Rule-Based Algorithm for Word Segmentation, In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL '97), pp321-328, Madrid, Spain, 1997

[3] J. Hockenmaier and C. Brew, Error-Driven Learning of Chinese Word Segmentation, Communications of COLIPS 8 (1), pp69-84, Singapore, 1998

[4] E. Brill, Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging, Computational Linguistics, vol. 21, no. 4, pp543-565, The MIT Press, USA, 1995

[5] E. Brill, Some advances in transformation-based part of speech tagging, In Proceedings of the Twelfth National Conference on Artificial Intelligence, pp722-727, Seattle, Washington, 1994

[6] E. Brill and P. Resnik, A rule-based approach to prepositional phrase attachment disambiguation, In Proceedings of the Fifteenth International Conference on Computational Linguistics (COLING-1994), pp998-1004, Tyoto, Japan, 1994

[7] E. Brill, Transformation-based error-driven parsing, In Proceedings of the Third International Workshop on Parsing Technologies, pp13-25, Tilburg, The Netherlands, 1993

[8] T. Yao, W. Ding and G. Erbach, Correcting word segmentation and part-of-speech tagging errors for Chinese named entity recognition, In Günter Hommel and Sheng Huanye (Eds.): The Internet Challenge: Technology and Applications, pp29-36, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002 (forthcoming)