Automated Grammar Checking of Tenses for ESL Writing

Nazlia Omar*, Nur Asma Mohd. Razali, and Saadiyah Darus

Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia, 43600, Bangi, Selangor, Malaysia
no@ftsm.ukm.my, nur_asma84@yahoo.com, adi@ukm.my

Abstract. Various word-processing system have been developed to identify grammatical errors and mark learners' essays. However, they are not specifically developed for Malaysian ESL (English as a second language) learners. A marking tool which is capable to identify errors in ESL writing for these learners is very much needed. Though there are numerous techniques adopted in grammar checking and automated essay marking system, research on the formation and use of heuristics to aid the construction of automated essay marking system has been scarce. This paper aims to introduce a heuristics based approach that can be utilized for grammar checking of tenses. This approach, which uses natural language processing technique, can be applied as part of the software requirement for a CBEM (Computer Based Essay Marking) system for ESL learners. The preliminary result based on the training set shows that the heuristics are useful and can improve the effectiveness of automated essay marking tool for detecting grammatical errors of tenses in ESL writing.

Keywords: Automated essay marking, natural language processing, computer-based essay marking, heuristics.

1 Introduction

English has become a language for communication worldwide and it has been enrolled as part of courses in university. Writing essays enables students to create avenues for their voices to be heard through communication particularly in writing. However, marking essays manually is an overwhelming task and time consuming for educators [1], especially when the size of the class increases. Larger classes reduce opportunities for learners' feedback on written work and learners in turn may have little contact with their lecturers. Darus et al. [2, 3, 4] shows that students prefer to receive feedback on errors in essay and their type, coherence of the text and organization of ideas.

In Malaysian educational context, English is used as a second language (ESL). ESL learners are those learners whose English is not their primary or first language (L1) but they are fairly able to write, read, speak or understand English. As learners go through various stages of acquisition of different elements of the second language,

* Corresponding author.

P. Wen et al. (Eds.): RSKT 2009, LNCS 5589, pp. 475–482, 2009.

they will make errors reflective of their second language acquisition processes. This paper aims to facilitate the task of essay marking especially to Malaysian ESL learners by automating the marking process. The task focuses on the identification of the grammatical errors in tenses made by the learners.

2 Related Work

Expert System for the Teaching of English [5] is a system which aims to provide English grammar learning. This system concentrates on checking grammatical errors in text files or individual sentences. The grammar files in the system are coded using BNF. In order to detect grammatical errors of sentences, a parser (internal code) is used. To obtain this internal code, a parser generator is used to derive the BNF coded grammar rules. The system consists of two kinds of environments; the teachers' environments and students' environments. Teachers are allowed to 'teach' the system and create a set of grammar files. In students' environments, the system will be interacting with the strings of words or sentences and display error messages if grammatical errors are detected. Facilities like editing text files, creating new ones on-line and option to store error file for future reference are also provided.

EasyEnglish [6] is an authoring tool applied to produce simple English writing, made suggestions for rephrasing and can be used as grammar checker. This tool is a part of IBM's internal SGML editing environment, Information Workbench. It is claimed that EasyEnglish is capable of detecting structural ambiguity in sentences. EasyEnglish is based on a full parse by English Slot Grammar and dictionary. User is allowed to specify any dictionary. The dictionary supports vocabulary checks for restricted word, identify acronyms or abbreviation in the text and give option to user to specify a controlled vocabulary. The tool is concentrated on grammar checking for syntactic, lexical and punctuation problems. EasyEnglish successfully detect lack of parallelism in coordination and list elements, passives, double negatives, long sentences, incomplete sentences, wrong pronoun case and long noun strings.

Tschichold et al. [7] proposed A New Grammar Checker for English. The prototype aimed for French native speakers writing in English. Three main components of this tool are a set of writing aids, a problem highlighter and a grammar checker. The writing aids included two monolingual and a bilingual dictionary, a verb conjugator, a small translating module and a comprehensive on-line grammar. Word highlighter allowed user to view all problematic words offers help such explanations and examples for each word. In the grammar checker, island processing is used in the process of detecting grammar errors. In island processing, the first step is to identify simple noun phrases. The second step is to identify the verb group and assign tense, voice and aspect features. The last step is error detection plus interaction with user. There are situations where the user is asked for additional information on the problem such as two possible corrections need to be chosen.

Although various tools have been proposed to facilitate the evaluation of student essays, they are not specifically developed for Malaysian ESL learners. A marking tool which is specifically developed to analyze errors in ESL writing is very much needed. In addition, research on the formation and use of heuristics to detect grammatical errors as part of the construction of automated essay marking system has been

scarce. This paper proposes new heuristics that can be used to mark essays automatically and detect grammatical errors in tenses. These techniques can be applied as part of the software requirement for a CBEM system for ESL learners.

3 Grammar Checking of Tenses

The development of this system is an extension of work carried out by Darus et al.[8]. An analysis of 400 essays of ESL learners at Universiti Kebangsaan Malaysia (UKM) for grammatical errors reveal that the most common error was tenses [9]. Based on the result of the analysis, the techniques and algorithm for detecting and analyzing tense errors in learners' essay are developed. The automated marking tool architecture mainly consists of two stages: the natural language stage and the logical stage. Figure 1 illustrates the processes involved in the automated marking tool for tenses in ESL Writing.

The natural language stage comprises of parsing the natural language input (in the form of tagged essays) into the system. The process begins by reading a plain input text file containing sentences written in English. For this purpose, a parser is used to parse the English sentences to obtain their part-of-speech (POS) tags before further processing. The parser used is CST's Part Of Speech Tagger [10]. The parsed text is then fed into the system to detect errors in tense. The final result consists of errors in tenses from the natural language input. The final result is displayed with a friendly user interface can be used as a feedback for students to improve the quality of their writing.

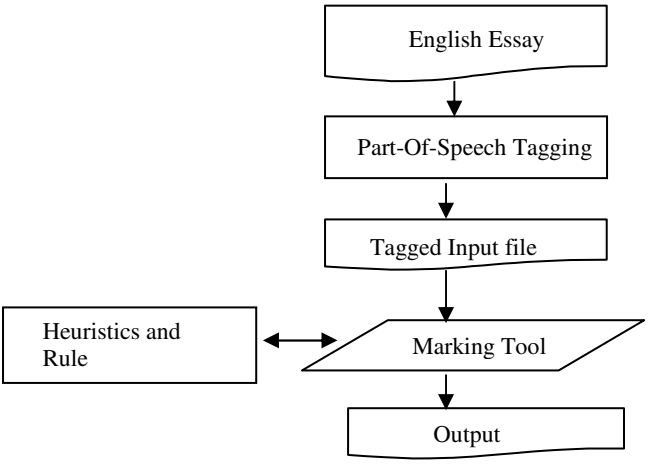

Fig. 1. Process in automated marking tool for ESL writing

3.1 Heuristics

Heuristics represent an indefinite assumption [11] often guided by common sense, to provide good but not necessarily optimal solutions to difficult problems, easily and quickly [12]. Through this research, a set of heuristics has been developed based on a corpus of ESL essays to detect grammatical errors in tenses. A complete set of the

heuristics can be found in [13]. Some examples in terms of sentences are provided [14] to illustrate the application of the heuristics which are based on syntax and context dependent. Some examples of heuristics are as follows:
Heuristic to identify error in Simple Future Tense:
If noun or pronoun is in the set of heuristic Future Tense followed by modal, check the verb after modal. If verb is tagged as VB, the sentence has no grammatical error.

Example sentence : Aziz will go to the market.
Tagged sentence : Aziz/NNP will/MD go/VB to/TO the/DT market/NN.

Heuristic to identify error in Present Progressive Tense:
If noun or pronoun is in the set of heuristic Present Progressive followed by 'is', check verb after 'is'. If verb is tagged as VBG, the sentence has no grammatical error.

Example sentence : Mary is reading the newspaper.
Tagged sentence : Mary/NNP is/VBZ reading/VBG the/DT newspaper/NN.

Heuristics to identify error in Present Perfect Progressive:
If noun or pronoun is in the set of heuristic Present Perfect Progressive followed by 'have been', check verb after 'have been'. If verb is tagged as VBG, the sentence has no grammatical error.

Example sentence : Your brothers have been waiting since 3.00 p.m.
Tagged sentence : Your/PRP brothers/NNS have/VBP been/VBN waiting/VBG since/IN 3.00/CD p.m/NN.

3.2 Algorithm

Figure 2 presents a portion of the algorithm to apply the heuristics to detect errors in Simple Present Tense. Figure 3 is a part of the algorithm to apply a selection of heuristics to detect errors in Simple Past Tense. In this algorithm, 'check' refers firstly to the identification of a match, which meets the heuristic's condition. Heuristics are applied to any relevant words in the sentences that meet the heuristics' criteria.

```
1. FOR each sentence in parsed and tagged text file, DO;
2. FOR each sentence, DO;
  2.1 IF word is of type 'noun' DO;
        IF noun is in the set of heuristic Simple Present 1,
          IF noun is tagged as NNS, check the immediate verb after the noun.
            IF verb is tagged as VBP or VB, print the following sentence: The
            sentence has no grammatical error.
          IF noun is tagged as PRPplural, check the immediate verb after the noun.
            IF verb is tagged as VBP or VB, print the following sentence: The sentence
            has no grammatical error.
        IF noun is in the set of heuristic Simple Present 2,
          IF noun is tagged as NN, check the immediate verb after the noun.
            IF verb is tagged as VBZ, print the following sentence: The sentence has no
            grammatical error.
```

Fig. 2. Extract from the algorithm for heuristics to detect error in Simple Present Tense

```
3. FOR each sentence, DO,
   3.1 FOR each noun, DO,
       IF noun is in the set of heuristic Past Tense 1,
         IF noun is tagged as NN, check the immediate verb.
           IF verb is tagged as VBD, print the following sentence: The sentence has no
           grammatical error.
         IF noun is tagged as NNP, check the immediate verb.
           IF verb is tagged as VBD, print the following sentence: The sentence has no
           grammatical error.
```

Fig. 3. Extract from the algorithm for heuristics to detect error in Simple Past Tense

To illustrate the process in automated marking tool, consider the following scenario 'Pollution':

"Nowadays, we always read, heard or watch about air pollution in the newspaper, radio or television. This situation showed that the air pollution is become serious day after day and effective ways must to be taken by the authorized sides to make sure this situation can be solved. Second, we also polluted our air by using many vehicles on the road and built more industrial park everyday. In addition, we also built many factories and most of these factories produce a lot of chemical waste including dangerous gasses."

The processing in the automated marking tool starts by reading the first line. From the first sentence, the system identifies the pronoun 'we' followed by the verb 'read', 'heard' and 'watch'. The marking tool identifies this sentence as having a grammatical error at the word 'heard' due to the application of the heuristic to identify error in Simple Present Tense 1. In the second sentence, the marking tool identifies the verb 'showed' that does not match with the verb tense in the first sentence. The marking tool identifies this sentence as having grammatical error at the word 'showed'. The result of all sentences will not be finalized until all lines have been read. Figure 4 shows the result from automated marking tool for scenario 'pollution'.

```
Nowadays , we always read , heard or watch about air
pollution in the newspaper , radio or
television




-->This sentence has grammatical error at word : heard
##
This situation showed that the air pollution is become
serious day after day and effective ways must to be taken
by the authorized sides to make sure this situation can be
solved
```

Fig. 4. Output from the marking tool

Table 3 shows a comparison of the result generated by the automated marking tool and the answer key, a solution produced by human analyst. In this context, answer keys are provided by linguists who have over ten years of experience in teaching English in a higher education institution. The automated marking tool has correctly detected most of the grammatical errors of tenses in the test dataset. However, there is only one error that has been detected incorrectly.

Table 3. Comparison of results obtained by automated marking tool and answer key

Sentence	*Error detected by automated marking tool*	*Answer Key*
1st	heard	heard
2nd	showed become authorized	showed become
3rd	polluted	polluted
4th	built	built

4 Experimental Result

An initial testing has been conducted in order to evaluate the results of the marking tool. Three possible outcomes in the evaluation are considered which are errors correctly detected, errors incorrectly detected, and errors undetected. The approach in this evaluation uses methods for evaluation Information Extraction systems evaluations i.e recall and precision [15]. Recall is percentage of all the possible correct answers produced by the system. Precision is the percentage of answers that are correctly identified by the system. In the context of this research, the definition of recall and precision are as follows:

4.1 Recall

Recall is the measure of the percentage of information available that is actually found. In this research context, it refers to the amount of the correct information returned by the system. The correct information is then compared with those produced by human analysts or answer keys. The following formula is used to calculate recall:

$$\text{Recall} = \frac{N_{correct}}{N_{key}}$$

The answer keys or N_{key} is actually the amount of correct information plus the number of undetected ones. Thus, the formula is refined as follows:

$$\text{Recall} = \frac{N_{correct}}{N_{correct} + N_{undetected}}$$

4.2 Precision

Precision is a measure of percentage of correctness of the information produced. It reflects the accuracy of the system in obtaining the correct result. The standard precision formula is as follows:

$$\text{Precision} = \frac{N_{correct}}{N_{correct} + N_{incorrect}}$$

The automated marking tool has been tested using a test dataset containing 50 essays. The system achieved a high average recall of 93.5%. In term of precision, the system scored an average of 78.8%. The results have shown consistent and encouraging results which indicate the potential use of this heuristics and rule based approach to practical automated marking tool. The heuristics and rule based approach are derived from the profiling errors of ESL learners' essays and observation of English grammar. The result support that this syntactic heuristics and rule based approach can be applied in the early stages of development CBEM (Computer Based Essay Marking) system for ESL learners.

5 Conclusion

We have described an approach for detecting grammatical errors of tenses in ESL writing. The formation of heuristics can be further developed to improve the accuracy of the result. For future work, weights may need to be assigned to each of the heuristic in order to support the level of confidence of each of the heuristics. The next step is to further evaluate the system using larger test dataset. Thus, further testing has been our main interest in the near future.

Acknowledgments. This study is funded by MOSTI (Ministry of Science, Technology and Innovation, Malaysia) with the following research code: e-ScienceFund 01-01-02-SF0092.

References

1. Page, E.B.: Computer grading of student prose – Using modern concept and software. Journal of Experimental Education 62, 127–142 (1994)
2. Darus, S., Hussin, S., Stapa, S.H.: Students expectations of a computer-based essay marking system. In: Mukundan, J. (ed.) Reflections, visions & dreams of practice: Selected papers from the IEC 2002 International Education Conference, pp. 197–204. ICT Learning Sdn Bhd, Kuala Lumpur (2001)

3. Darus, S.: A prospect of automatic essay marking. In: International Seminar on Language in the Global Context: Implications for the Language Classroom, RELC, Singapore (1999)
4. Darus, S., Stapa, S.H., Hussin, S., Koo, Y.L.: A Survey of Computer-Based Essay Marking (CBEM) Systems. In: Proceedings of the International Conference on Education and ICT in the New Millennium, pp. 519–529. Faculty of Educational Studies, Universiti Putra Malaysia (2000)
5. Chan, C.K.: An Expert System for the Teaching of English Grammar. In: Conference on Computer and Communication Systems, Hong Kong. IEEE Region, vol. 10, pp. 181–184 (1990)
6. Bernth, A.: EasyEnglish: A Tool for Improving Document Quality. In: Proceedings of the Fifth Conference on Applied Natural Language Processing, Association for Computational Lingusitic, pp. 159–165 (1997)
7. Tschichold, C., Bodmer, F., Cornu, E., Grosjean, F., Grosjean, L., Kubler, N., Lewy, N., Tschumi, C.: Developing a New Grammar Checker for English as a Second Language. In: Proceedings of the From Research to Search to Commercial Applications Workshops, pp. 7–12
8. Darus, S.: A framework of a CBEM for ESL writing at institutions of higher learning in Malaysia, PhD Thesis, Universiti Kebangsaan Malaysia (2005)
9. Darus., S., Mohd Maasum, T.N.R.T., Stapa, S.H., Omar, N., Aziz, M.J.A.: Developing an error analysis marking tool for ESL learners. In: 7th WSEAS International Conference on Applied Computer Science, pp. 356–359 (2007)
10. CST's Part-Of-Speech tagger, `http://cst.dk/online/pos_tagger/uk/index.html`
11. Tjoa, A.M., Berger, L.: Transformation of requirements Specifications Expressed in Natural Language into an EER Model. In: Proceeding of the 12th International Conference on Approach, Airlington, Texas, USA, pp. 206–217 (1993)
12. Zanakis, S.H., Evans, J.R.: Heuristics 'Optimization': Why, When and How to use it. Interfaces, 84–91 (1981)
13. Mohd Razali, N.A., Omar, N., Darus, S., Mohd Maasum, T.N.R.T., Ab. Aziz, M.J.: Pendekatan Heuristik dan Peraturan untuk Pemeriksaan Nahu secara Automatik bagi Penulisan Esei Bahasa Inggeris sebagai Bahasa Kedua. In: Prosiding Persidangan Kebangsaan Sains Pengaturcaraan Atur 2007, Malay. Pan Pacific KLIA, pp. 77–88 (2007)
14. Mohd Razali, N.A., Omar, N., Darus, S.: Heuristics and Rule-based Approach for Automated Marking Tool for ESL Writing. In: Proceedings of the International Symposium on Information Technology 2008 (ITSIM 2008). Kuala Lumpur Convention Centre, Malaysia, August 26-29, 2008, pp. 144–149 (2008) IEEE Cat. No. CFP0833E-PRT, ISBN 978-1-4244-2327-9
15. Jurafsky, D., Martin, J.H.: Speech and Language Processing; An Introduction to Natural Language Processing. In: Computational Linguistics and Speech Recognition, Prentice Hall, Inc., Upper Saddle River (2000)