http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 9 (3): 605-608, 2010
ISSN 1812-5638

Automatic Chinglish Identification Based on Semantic Distances Calculation

[1]Wen Zhuge and [2]Jingyu Hua
[1]College of Foreign Language, Hangzhou Dianzi University, Hangzhou, 310023, China
[2]College of Information Engineering, Zhejiang University of Technology, Hangzhou, 310032, China

Abstract: Writing has been considered as an effective way to measure a language learner's language proficiency. With labor and resources saved, writing test is entering into an era in need of Automated Essay Scoring (AES). However, its further promotion in China is limited by the negative transfer of non-native learners. Therefore, this study proposed a new way to identify Chinglish, which obstructs the development of AES in China. This WordNet-based method starts with dealing with semantic relations between English verbs, calculates semantic distances between subjects and objects and then realizes the identification of Chinglish by threshold. Experiments conducted in one university show that the proposed way performs well in identifying Chinglish in college students' English essays.

Key words: Automated essay scoring, Chinglish identification, semantic calculation, WordNet

INTRODUCTION

Writing has been considered as an effective method to measure a language learner's language proficiency. It accounts for a large proportion of classroom teaching as well as proficiency tests. However, the traditional way to score an essay not only costs a great deal of labor and material resources, but also is greatly affected by scorers' language ability and personal preferences, therefore, reducing the credibility of the test (Wen, 2003). However, Automated Essay Scoring (AES) based on corpus and artificial intelligence technology could perform this task with high efficiency and impersonality (Streeter *et al*., 2006). With labor and resources saved, language testing is entering into an era in need of AES systems.

Currently, there already exist an ocean of AES systems overseas, such as PEG, IEA and Intelli-Metric, which are based on shallow parsing, latent semantic analysis and the combination of artificial intelligence and statistics, respectively. They have already achieved a relatively accurate scoring task by Ge and Chen (2007), Hearst *et al*. (2000), Elliot (2003) and Landauer *et al*. (2003). However, these systems are English speakers oriented, thus they neglect the problems caused by the negative transfer of non-native learners. As far as Chinese learners are concerned, Chinglish sentences i.e., the hybrid of Chinese way of thinking and English syntax, should be given priority in the research of AES technologies.

Since, Chinglish sentences accord with English syntax, it is difficult for the common syntactic analyzers employed by AES system to identify these special linguistic phenomena. Utilizing the semantic web put forward by WordNet, the researchers start with the semantic relations between English verbs, deal with semantic relations between English verbs, calculate semantic distances between subjects and objects and then realize the identification of Chinglish by threshold in this study.

CHINGLISH IN ESSAYS

Definition of Chinglish: Transfer is a term used by psychologists in their accounts of the way in which present learning is affected by past learning (Miao and Ni, 2000). Odlin has suggested that transfer is the influence resulting from the similarities and differences between the target language and any other language that has been previously (and perhaps imperfectly) acquired (Odlin, 2001). When learning a foreign language, an individual already knows his mother tongue, so the knowledge and skill of the native language will be transferred to the foreign language unconsciously. Since, the way one thinks determines the way one speaks, the differences between the native and the foreign language in linguistic structure, social culture and logic thought, predestine the existence of negative transfer in second language learning.

The Han nationality puts more emphasis on the whole when they think, which, in turn results in an analytical language Chinese, while British and Americans, who weigh reason and analysis more, create a synthetic language English. This difference in way of thinking directly leads to the appearance of Chinglish.

Corresponding Author: Wen Zhuge, College of Foreign Language, Hangzhou Dianzi University, Hangzhou, 310023, China

Table 1: Comparison of essay scores given by teachers and e-rater

Mother tongue	Teachers	E-Rater
Arabic	3.83±0.973	3.67±0.947
Chinese	4.09±0.884	4.12±1
Spanish	3.96±0.986	3.70±0.915
American	4.96±0.624	4.93±0.814

Data is expressed as average±SD

Generally speaking, Chinglish appears in two forms: vocabulary and theme (Zhao and Liu, 2006). The former is caused by the unequivalence of the concept aroused by certain words in Chinese and English, whose meanings are simply taken for granted in one culture. Take the sentence, "She quarreled with her boyfriend and ran out in the big rain" for example. In Chinese, the concept of big covers almost every aspect in our life; yet, the counterpart of this concept in English varies. They have strong wind, heavy rain, loud sound' and wide road etc. The latter results from the confusion of a subject and a theme. A subject in an English sentence is usually a person or a thing, represented by a noun, pronoun, or noun phrase. But subject can take any form in a Chinese sentence. Any language elements can appear in the place of a subject, such as The table has four legs.

Obstacles caused in AES: As mentioned above, Chinglish sentences are correct in terms of syntax, thus AES systems mainly based on syntax analysis will make wrong judgment about these sentences.

Burstein and Chodorow (1999) has made a comparison experiment with E-Rater, in which he collected essays from both native learners and nonnative learners and had these essays scored by teachers and E-Rater, respectively. His research results can be shown in Table 1.

From Table 1, we can clearly see that only essays from Chinese learners receive higher marks from E-Rater than teachers. To account for it, Ge and Chen (2007) pointed that Chinese learners' large vocabulary shadowed their misuse of sentence structure, thus cheating the AES system and affecting its accuracy.

AUTOMATIC CHINGLISH IDENTIFICATION

Semantic relations in WordNet: Being an electronic dictionary based on psychological research, WordNet is an effective combination of both traditional lexicographical knowledge and modern computer technology. It employs the spelling system, which is familiar to anyone with some knowledge of English, to symbolize word form and introduces synsets to stand for word sense.

The most ambitious feature of WordNet, however, is its attempt to organize lexical information in terms of word senses, rather than word forms. In that respect, WordNet resembles a thesaurus more than a dictionary. Its intricate yet quiet clear representation of hyponymy, which is realized by pointers, chains and lists in its database, created a hierarchical semantic system, or an inheritance system for words. Thus, one can easily tracethe hyperym, hyponym, co-hyponym and even holonym and meronym of certain word easily through WordNet browser.

Table 2: Relation between nouns and verbs

Vocabulary	Part of speech	Semantic category	Agent	Patient
Table	n.	Object	Human	None
Leg	n.	Entity	Entity	None
Father	n.	Human	Entity	Entity
Wash	v.	Change	Human, object	Object
Have	v.	Possession	Living thing	Entity
Big	adj.	Size	None	None

All the nouns in WordNet form a single thematic hierarchical tree, whose root is entity, because inheritance underlines their semantic relations. Adjectives and adverbs assemble satellite synsets of antonyms, since the bipolar nature is their unique and outstanding feature. Altogether, there are 25 categories in WordNet, which can be furthered grouped into 11 classes: entity, abstraction, psycho features, natural phenomenon, activity, event, group, location, possession, shape and state. Similarly, verbs, which scattered in 15 semantic domains, weave a network of entailment, for this network covers almost all the semantic relations among verbs (Fellbaum, 1998).

Semantic description of the new method: There is always a semantic database behind an AES system, the scheme of which directly affects the credibility of the system. It will be beneficial to the identification of Chinglish, if we add collocation information, along with the description of semantic features possessed by the subject and object of a certain verb on the platform set up by WordNet. For instance, to tag one semantic entry of the verb heavy (unusually great in degree or quantity or number), the domain of the noun it describes should be included.

According to Case Grammar, arguments covering agent, patient, location and instrument/benefactive etc. are all attached to verbs. For most verbs, subject, object and benefactive are what researchers care most. Thus, we exclude the other semantic information in our scheme and depict a web between nouns and verbs in Table 2.

Take wash for example, it is marked as a change verb in our scheme, whose agent should be human beings or machines, who are capable of washing. Besides, wash in English is an intransitive verb, so a patient other than location cannot follow it. Thus, I washed in the dog is incorrect while I washed in the house is acceptable. This scheme is quite effective in identifying typical Chinglish sentences.

Because of the limited space of this study, the way to construct such a semantic database will not be mentioned. All the analysis below will be performed with a database constructed by the semantic dictionary of English verbs based on WordNet and FrameNet.

Identification of Chinglish: With the help of the noun categories, as well as the semantic web of hypernym and hyponym provided by WordNet, this new method starts with semantic relations between verbs. By calculating semantic distances between subjects and objects, the identification of Chinglish can be realized by threshold. Take the two sentences mentioned in the precious section as examples; we can employ the method to make a judgment.

- **Example I:** The table has four legs

where, the syntactic analyzer tells us that has is the verb in this sentence. Matching it with its stem have in the database, we got the semantic restriction of its agent and patient, being living thing and entity, respectively. Then, the method casts back the semantic categories of the agent and patient in the current sentence i.e., table and leg. Finally, WordNet returns the research as:

- 04379964 06 n 01 Table 2, 00203405725 n 0000 ~ 03201208 n 0000 | a piece of furniture with tableware for a meal laid out on it; I reserved a table at my favorite restaurant

As we've mentioned earlier that means is the hyponym of, we can see that synset {home, place} has a hypernym, numbered 08558963 i.e., synset {residence, abode}. As long as does not point to Ø, the search keeps going. Ultimately, we've got:

Table furniture, piece of furniture, article of furniture

- Furnishing
- Instrumentality, instrumentation
- Artifact, artefact
- Whole, unit
- Object, physical object
- Physical entity
- Entity

The subject locates under the root nod of {physical entity}, while according to the semantic description of the subject features of have stored in the database, its subject should be on the tree under the nod {living thing}. Both {physical entity} and {living-thing} are on the second layer of a semantic tree. According to ontology, the higher two nods are, the remoter they will become. So the subject home and the verb have do not match semantically. That is to say, although this sentence is grammatically correct, it is improper semantically.

- **Example II:** She quarreled with her boyfriend and ran out in the big rain

Because big is adjective here, we can make use of the noun it describes, rain, to perform the calculation. After calculation, we get:

Big is a value of size (the physical magnitude of something); rain, rainfall precipitation, downfall

- Weather, weather condition, atmospheric condition
- Atmospheric phenomenon
- Physical phenomenon
- Natural phenomenon
- Phenomenon
- Process, physical process
- Physical entity
- Entity

The comparison between the semantic categories where rain and the noun which big describes i.e., {process} and {physical magnitude}, tells us that these two nods also belong to two different tress and therefore, Chinglish can be found here. Thus, it can be seen that, the semantic categories and the semantic relations provided by WordNet can be effectively utilized to identify Chinglish in an essay. Generally, we can analyze the structure of sentences by syntax analyzer, then match the semantic description of the collocation information of the verb with its subjects or objects, finally calculate the semantic distances to make a decision. The whole process can be summarized in Fig. 1.

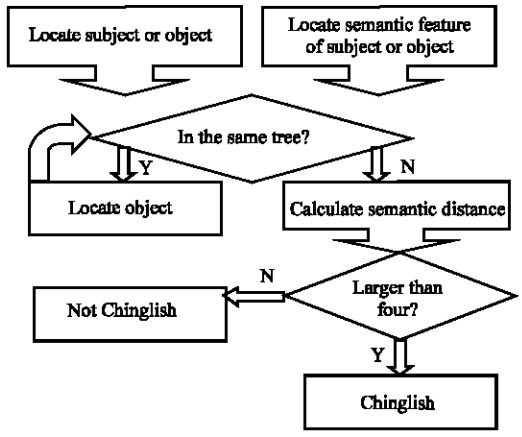

Fig.1: The process of identifying Chinglish

TESTS AND RESULTS

Besides the above examples, we also made experiments with undergraduates in Zhejiang province in China, where three hundred randomly-collected sample essays written by freshman majored in engineering sciences are collected. Among these essays, there are 147 Chinglish sentences identified by English teachers and our method successfully identified 75% of them, which approves its efficiency. In detail, we find that these sentences can be grouped into three. The common problem in group one, with 41 sentences in it, lies in the syntactic level i.e., there is either none verbs at all, or several verbs in a single sentence, such as we against it, we run over meet her and so on. Although, the proposed method only recognized 23 of them, especially the ones with double or triple verbs, this disadvantage can be easily compensated with the help of a syntactic analyzer, which specialized in detecting errors in sentence structure. Group two, covering the majority of these sentences with up to one hundred and three sentences, involves misuse of words. This may caused either by the incorrect choices of parts of speech of certain words, or the simple mapping of a concept in Chinese into English, in instance, we worked hardly. It is in this group that the proposed method gained a high accuracy rate, 88% actually, for the wrongly chosen words usually fall into a totally different semantic domain listed in our database from that of its agent or client. The last group, containing twenty one sentences like American Chinese not enough, or walk past no miss past, is a salmagundi, which is hard to understand, but easy to identify, thus effective again for our method. From above discussions, we definitely observe the efficiency of our method and believe the proposed method must be beneficial for oriental AES systems.

CONCLUSIONS

This study proposed a novel automatic Chinglish identifier based on WordNet and semantic information description. Both the examples and the test results produce high identification rate, hence, our method can identify Chinglish effectively. Moreover, this new method can be used as a supplement to the existing AES systems e.g., merging the structure of FrameNet and matching its semantic concept with HowNet, by which the power of this method could be strengthened and help improve the accuracy of existing AES systems in the future.

ACKNOWLEDGMENT

This study is supported by Zhejiang provincial NSF project (Y1090645) and the open research fund of Zhejiang Key Lab. of Opt. Commun., Zhejiang University of Technology.

REFERENCES

Burstein, J. and M. Chodorow, 1999. Automated essay scoring for nonnative English speakers. Proceedings Joint Symposium of the Association of Computational Linguistics and the International Association of Language Learning Technologies, June 1999, College Park, Maryland, pp: 68-75.

Elliot, S., 2003. Automated Essay Scoring: A Cross Disciplinary Perspective. LEA, New Jercey, pp: 71-86.

Fellbaum, C., 1998. WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA., USA., ISBN: 0-262-06197-X.

Ge, S.L. and X.X. Chen, 2007. An overview of current automated essay scoring techniques. Media in Foreign Language Instruction} (China), (117): 25-29, 2007. doi: CNKI:SUN:WYDH.0.2007-05-006. http://caod.oriprobe.com/articles/472525/An_Overview_of_Current_Automated_Essay_Scoring_Techniques.htm.

Hearst, M., K. Kukich and L. Hirschman, 2000. The debate on automated essay grading. IEEE Intel. Syst., 15: 22-37.

Landauer, T., D. Laham and P. Foltz, 2003. Automated Essay Scoring and Annotation of Essays with the Intelligent Essay Assessor. In: Automated Essay Scoring: A Cross Disciplinary Perspective, Shermis, M.D. and J. Burstein (Eds.). LEA, New Jercey, pp: 87-112.

Miao, J. and X.G. Ni, 2000. A theoretical study of language transfer. J. Inner Mongolia Coll. Educ., 13: 12-17.

Odlin, T., 2001. Language Transfer: Cross-Linguistic Influence in Language Learning. Shanghai Foreign Language Education Press, Shanghai.

Streeter, L., J. Pstoka, D. Laham and D. MacCuish, 2006. The credible grading machine: Automated essay scoring in the dod. http://www.pearsonkt.com/papers/CredGrading2002.pdf.

Wen, J.F., 2003. A reliable scoring system for norm-referenced tests of English writing. J. Guangzhou Univ., 2: 84-87.

Zhao, Y.S. and N.N. Liu, 2006. Error analysis about college English writing based on the difference between Chinese thinking and English thinking. J. Weifang Educ. Coll., 21: 21-24.