An Expertise-based Vocabulary Suggestion System: A Case Study of LEXiTRON

Kanokorn Trakultaweekoon[1], Wasan Na Chai[1], and Thepchai Supnithi[1]

[1] Human Language Technology Laboratory, National Electronic and Computer Technology Center, Thailand Science Park, Pathumthani, 12120, Thailand
{kanokorn.trakultaweekoon, wasan.na_chai, thepchai.supnithi}@nectec.or.th

Abstract. This paper presents an expertise-based vocabulary suggestion system for community based online dictionary. The distinguished point of this paper is that we utilize confidence scores of suggesting and voting activities to reduce linguists' burden of validation. The scores, which imply the expert level, are calculated by the numbers of accepted and unaccepted vocabularies that the user suggested or voted. All suggested vocabularies are calculated based on confidence scores of their suggesters and voters. The unaccepted vocabularies from the community members are filtered out from the validation list.

1 Introduction

At present, an online dictionary plays an essential role as a tool for roughly understanding and studying foreign languages. Online dictionary construction and maintenance are nontrivial due to slow growth of dictionary size and linguists validation load.

A general problem of dictionary-online development is to update new vocabularies. Even though, experts are able to collect and store new vocabularies, the chosen words might not meet user's interests or they possibly are too particular. Therefore, it is easier to let users suggest and decide the words by themselves.

Some famous online dictionaries, for example Longdo [1] and LEXiTRON [2], provide an environment suggestion. However, this approach exhibits difficulty of maintenance because the system requires linguistics qualification from suggesters, in addition, approved linguist are burdened by a large amount of vocabulary validation.

Recently, expertise-based communities such as KUI [3] and Gotoknow [4] have gained popularity in terms of collaboration due to user expertise scoring. We were motivated by this approach because dictionary maintenance can be reckoned as an expertise-based community. Vocabulary suggestion can be promoted if the suggesters and voters are experts inferred from confidence scores.

In this paper, we propose an expertise-based vocabulary suggestion system. We reduce validation load of linguists by using confidence scores to imply the expertise of users.

The rest of the paper is represented as follows. Section 2 describes our vocabulary suggestion system architecture and its detail. Next, the processes of vocabulary suggestion, voting, and validation are expressed in Section 3. Section 4 discusses the

calculation of our confidence score. The vote mechanism and its criteria are explained in Section 5. In Section 6, the result of our system is shown. Finally, we sum up the paper and list up future work in Section 7.

2 Vocabulary Suggestion System Architecture

We improved the vote mechanism with confidence score based on users' vocabulary suggestions and votes. Our system illustrated in Figure 1, consists of four modules; suggestion module, vote mechanism module, validation module, and update module.

- **Suggestion Module**

The objective of this module is to support suggesters to suggest new vocabularies with details. This module has two components; new vocabulary module, and recommendation mechanism module. The new vocabulary module provides an interface to assist users to add new vocabularies. The recommendation mechanism provides a list of unknown vocabularies for suggesters to add new vocabularies that meet user's interests.

- **Vote Mechanism Module**

This module reduces validation load of linguists. It composes of three components. The first one is collectable voting, a function to add voting score. The second is visualization, a graphical user interface of scores in each acceptance level. The last one is vote acceptation, the criteria to filter improper words before linguists correct and add them into dictionary.

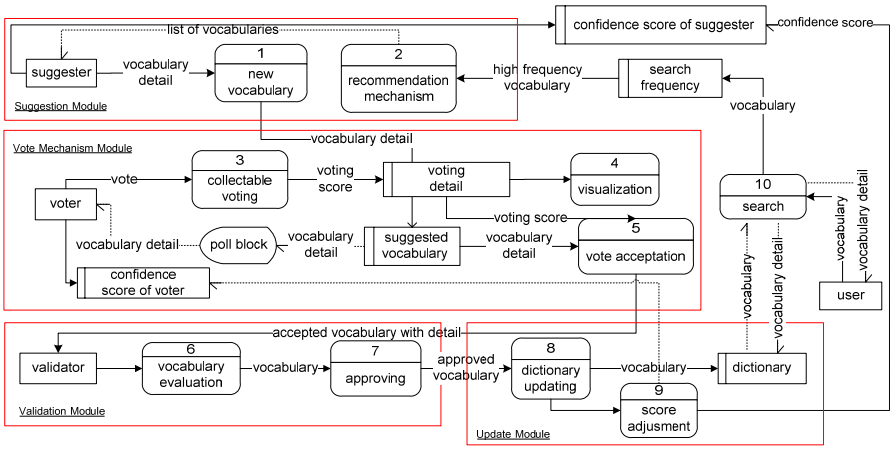

Fig. 1. An overview of the system

- **Validation Module**

There are two steps of validating accepted vocabularies. First, linguists inspect details of the vocabularies. If they accept, the vocabularies will be approved and sent to update module.

- **Update Module**

There are two components in this module, score adjustment and dictionary updating. The former is to update confidence score according to the reliability of each user. The latter is to update vocabularies. Approved vocabularies will be added into dictionary. In the contrary, improper word, for example impolite word, will be deleted. The confidence scores of users will be recalculated with respect to the number of their approved and disapproved vocabularies. The formula will be described in Section 4.

3 Vocabulary Suggestion, Voting, and Validation Process

The process of the system can be explained as follows. There are three alternatives for users to add new vocabularies. The first alternative is to suggest them directly. The second alternative is to suggest them when the users look up for unknown word. The last alternative is to use the recommendation system. The suggester is required to provide primitive information, such as vocabulary label, part-of-speech, or meaning to check whether this word exists in, dictionary or not. The non-existing vocabularies will be inserted into dictionary.

After the user suggested them, it will be queued up for voting. Each user has only one voting right for any vocabulary except his/her own suggested word. There are three voting methods. Firstly, he/she can vote the vocabulary automatically shown in the pop-up windows when he/she looks up any vocabulary. Secondly, the voter can vote directly via a vote link. Lastly, the voters can vote for the vocabulary that is shown in poll box. The suggested vocabulary will be randomly displayed weekly. The voting score will be stored in database when user votes for the vocabulary. The score in each acceptance level is visualized in graph.

System will roughly check whether any vocabularies are accepted or not. The accepted ones will be transferred to linguists validated environment. Otherwise, it will be deleted. The linguists will verify and correct those words

Once validated words are added into dictionary, the confidence score will be increased. On the other hand, improper words will be deleted, and the score will be decreased. After above tasks are done, the system will update confidence score of both suggesters and voters who contribute the word.

4 Confidence Score

Confidence score implies the user's expert level. Every user has different role scores (Suggester, Voter) that depend on the expertise in each activity. The confidence level can be divided as follows.

1) *Strong Accept*: the given details are correct and acceptable.
2) *Weak Accept*: the given details are slightly incorrect and should be modified.
3) *Delete*: most voters are unsatisfied with the given vocabulary.

4.1 Confidence score of suggester

This score is calculated by the number of accepted and unaccepted vocabularies a user suggested. The confidence formula is defined below.

$$Conf_S = \left(\frac{nCorrect_S}{n_S}\right) - \left(\frac{nIncorrect_S}{n_S}\right)^{\frac{N_S}{n_S}} \quad \textbf{(1)}$$

Where,
$Conf_S$: confidence score of suggester
$nCorrect_S$: the number of suggested vocabularies that is "*Accept*" level
$nIncorrect_S$: the number of suggested vocabularies that is "*Delete*" level
n_S: the number of vocabularies that are suggested by oneself
N_S: the number of vocabularies that are suggested by all suggesters

4.2 Confidence score of voter

This score is calculated by the number of accepted and unaccepted vocabularies a user votes for. The confidence formula is defined as follows.

$$Conf_V = \left(\frac{nCorrect_{V(x)}}{n_{V(x)}}\right) - \left(\frac{nIncorrect_{V(x)}}{n_{V(x)}}\right)^{\frac{N_{V(x)}}{n_{V(x)}}} \quad \textbf{(2)}$$

Where,
$Conf_V$: confidence score of voter
$nCorrect_V$: the number of vocabularies that linguist and voter assign same level
$nIncorrect_V$: the number of vocabularies that linguist and voter assign different level
n_V: the number of vocabularies that are voted by oneself
N_V: the number of vocabularies that are voted by all voters
(x): levels of acceptance score are *Strong Accept*, *Weak Accept*, and *Delete* level

The above formulas imply the effect of incorrect suggesting and voting to confidence score. Confidence score of suggester/voter will be increased when vocabularies are accepted. Otherwise, the confidence score is decreased. Confidence score ranges from -1 to 1 according to the expertise of each user.

5 Vote Mechanism

This mechanism improves vocabulary suggestion system by filtering vocabularies unaccepted by the community. If a high confident suggester suggests new words, they are tentatively accepted. This performs time reduction on dictionary improvement.

Figure 2 shows the pseudo-code of criteria used for assigning status for each suggested word.

The acceptance score will be applied to decide the final result of word acceptance status. For example, if the word level is *Strong Accept* and not *Delete*, it will be assigned as **SA** status. If the word level is not *Weak Accept* and *Delete*, it will be assigned as **D** status. Otherwise, it will be assigned as **WA** status. The filtrated vocabulary criteria are shown in Table 1.

In our vote acceptation process, only words in **SA** and **WA** result will be sent to validation module. The word in **D** result will be automatically deleted. Once the acceptance processes are done, confidence value of involving users will be adjusted.

```
Input
    strong-accept score (Ss) = score of vote in strong accept
level
    weak-accept score (Sw) = score of vote in weak accept level
    delete score (Sd) = score of vote in delete level
Output
    Word status i.e. Strong_Accept, Weak_Accept and Delete
Initial
    Strong_Accept = false
    Weak_Accept = false
    Delete = false
total score (St) = Ss + Sw + Sd
IF ( Ss > ((1-Conf_S) * St) )  THEN
    Strong_Accept = true
ELSE IF ( (Ss+Sw) > ((1-Conf_S) * St) ) THEN
    Weak_Accept = true
END IF
IF ( Sd > (Conf_S * St) ) THEN
    Delete = true
END IF
```

Fig. 2. Pseudo code of acceptance criteria

Table 1. Filtered vocabulary table

Vocab_status			**Result**
Strong Accept	**Weak Accept**	**Delete**	
O		O	WA
O		X	SA
	O	O	WA
	O	X	WA
	X	O	D
	X	X	WA

O=Yes, X=No
SA=Strong Accept,
WA=Weak Accept,
D=Delete

6 Result

The vote mechanism has already been released since January, 2007. However, the confidence score was plugged in recently on February, 2008. Currently, we obtained the following significant statistics.

Table 2. Statistics of validated vocabulary

Vocabulary Status	**Voting**	**Validating**	
		Accept	**Reject**
SA	38 23.75%	38 100%	0 0%
WA	122 76.25%	78 63.93%	44 36.07%
total	**160**	**116 72.5%**	**44 27.5 %**

- There are 3,010 items suggested by 473 suggesters.
- There are 2,289 items voted by 10,303 voters.
- There are 73 improper words automatically filtered by voting.

Table 2 shows the statistics of the voted vocabularies which are accepted or rejected by linguists. The strong words accepted is 100% accepted by our linguists.

In fact, the efficiency of system

depends on expertise of voters and suggesters. However, all vocabularies will be audited by linguists

7 Conclusion

In this paper, we developed an expertise-based vocabulary suggestion system. The main objectives are to realize suggester's skill and to reduce a burden of linguists to check all vocabularies. If suggesters have high confidence score, their vocabularies can be tentatively acceptable. Unless we utilize confidence score for checking vocabularies in the system, linguists will validate laboriously.

Based on linguists' opinion, the system reduces work and time consumption. Linguists are possible to verify vocabularies referring to strong accept level and to trust a suggester who has a high confidence score.

Currently, there are 473 suggesters and 10,303 voters who join this expertise-based vocabulary suggestion system.

In the future work, we plan to analyze data to identify volunteers who are expert for voting or suggesting. Moreover, we currently concentrate on only suggesting and voting activities. It is possible to apply this concept into other viewpoints, for example, specific knowledge, gender, age, and so on, to gain various specific dictionaries such as, domain specific dictionary, gender –based dictionary, and teen-slang dictionary.

References

1. Longdo, http://www.longdo.com
2. LEXiTRON, http://lexitron.nectec.or.th
3. Charoenporn, T., Sornlertlamvanich, V., Robkop, K.: KUI: an ubiquitous tool for collective intelligence development. Proc. Of the IJCNLP-08 Workshop on NLP for Less Privileged Languages, Hyderabad, India, pp. 13--18 (2008)
4. Gotoknow, http://gotoknow.or.th
5. Trakultaweekoon, K., Porkaew, P., Supnithi, T.: LEXiTRON Vocabulary Suggestion System with Recommendation and Vote Mechanism. Proc. Of conference of SNLP- 2007, Chonburi, Thailand, pp. 43--48 (2007)
6. Mangeot, M., Sérasset, G.: Frameworks, implementation and open problems for the collaborative building of a multilingual lexical database. In Grace Ngai, Pascale Fung, and Kenneth W. Church,editors. Proc. Of SEMANET Workshop, Post COLING 2002 Workshop, Taipei, Taiwan, pp. 9--15 (2002)
7. Palingoon P., Chantanapraiwan, P., Theerawattanasuk, S., Charaoenporn, T., Sornlertlumvanich, V.: Qualitative and Quantitative Approaches in Bilingual Corpus-Based Dictionary. Proc. Of conference of SNLP-Oriental COCOSDA 2002, Prachuapkirikhan, Thailand, pp. 152--158 (2002)