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

Kanokorn Trakultaweekoon¹, Wasan Na Chai¹, Chaiyaporn Sukchareon¹,  
and Thepchai Supnithi¹

¹ Human Language Technology Laboratory, National Electronic and Computer Technology  
Center, Thailand Science Park, Pathumthani, 12120, Thailand  
{kanokorn.trakultaweekoon, wasan_na_chai, chaiyaporn.sukchareon,  
thepchai.supnithi}@nectec.or.th

Abstract. This paper presents an expertise-based vocabulary suggestion system  
for community based online dictionary maintenance. The distinguished point of  
this paper is that we utilize confidence scores of suggestion and vote activities  
to reduce linguists’ burden of validation. The scores, which imply the expert  
level, are calculated by the numbers of accepted and unaccepted vocabularies  
the user suggested or voted. All suggested vocabularies are promoted or  
penalized by 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.

The general problem of dictionary-online development is to update new  
vocabularies. Our online dictionary Lexitron also has such problem. Even though,  
exerts 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.

Most of online dictionaries are exploiting community-based maintenance in order  
of rapid vocabulary collection. Some famous online dictionaries, such as Longdo and  
Lexitron, 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 [T.Charoenporn,  
V.Sornlertlamvanich, 2008] and Gotoknow.org [http://gotoknow.org] have gained  
popularity in terms of collaboration due to user expertise scoring. We were motivated  
by this approach because dictionary maintenance can be seen 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 implying 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.

2   Vocabulary Suggestion System Architecture

We improved the vote mechanism module with confidence value of each user by determining 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.

2.1   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.
2.2 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.

2.3 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.

2.4 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 further described in details 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 from 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 vocabularies 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 vocabularies. 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 the words that are strong accepted or weak accepted.
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 score can be divided as follows.

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.

\[
ConfS = \frac{nCorrectS}{nS} - \frac{nIncorrectS}{nS}
\]

Where,
- ConfS: confidence score of suggester
- nCorrectS: the number of suggested vocabularies that linguist assigns “accept” status
- nIncorrectS: the number of suggested vocabularies that linguist assigns “delete” status
- nS: the number of vocabularies that are suggested by oneself
- NS: 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.

\[
ConfV = \frac{nCorrectV}{nV} - \frac{nIncorrectV}{nV}
\]

Where,
- ConfV: confidence score of voter
- nCorrectV: the number of voted vocabularies that linguist and voter assign same status
- nIncorrectV: the number of voted vocabularies that linguist and voter assign different status
- nV: the number of vocabularies that are voted by oneself
- NV: 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. A reliable user has better score and his vocabulary suggesting and voting tend to be more acceptable.
5 Vote Mechanism

There are three levels of acceptance score.
1) Strong Accept: the given word’s details are almost complete.
2) Weak Accept: the given word’s details are incomplete and should be modified.
3) Delete: most voters are unsatisfied with the given vocabulary.

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 status of word compared with filtrated vocabulary is shown in Table 1.

\[
\begin{align*}
\text{strong-accept score (Ss)} &= \text{score of vote in strong accept level} \\
\text{weak-accept score (Sw)} &= \text{score of vote in weak accept level} \\
\text{delete score (Sd)} &= \text{score of vote in delete level} \\
\text{total score (St)} &= \text{Ss} + \text{Sw} + \text{Sd} \\
\text{IF}( \text{Ss} > ((1-\text{ConfS}) \cdot \text{St}) ) \text{ THEN } \text{Strong Accept} = \text{true} \\
\text{ELSE IF}( \text{Ss} + \text{Sw} > ((1-\text{ConfS}) \cdot \text{St}) ) \text{ THEN } \text{Weak Accept} = \text{true} \\
\text{ELSE IF}( \text{Sd} > (\text{ConfS} \cdot \text{St}) ) \text{ THEN } \text{Delete} = \text{true} \\
\text{END IF}
\end{align*}
\]

Fig. 2. Pseudo code of acceptance criteria

The score of strong accept, weak accept and delete will be considered in order to assign status to compare with criteria of acceptance word. If the word status are strong accept and not delete, it will be SA status. If not, it is delete status only, it will be D status. Otherwise, it will be WA status.

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 deleted. When the acceptance process are finished, confidence value will be adjusted.

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.
- There are 2,820 items suggested by 452 suggesters.
- There are 2,114 items voted by 9,826 voters.
- There are 42 improper words automatically filtered by voting.

Table 2 shows the statistics of the voted vocabularies which are accepted or rejected by linguists. The words are strong accepted tend to be more acceptable.

In fact, the efficiency of system depends on expertise of voters and suggesters. However, all vocabularies will be audited by linguists.
### Table 1. Filtered vocabulary table

<table>
<thead>
<tr>
<th>vocab status</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong Accept</td>
<td></td>
</tr>
<tr>
<td>Weak Accept</td>
<td></td>
</tr>
<tr>
<td>Delete</td>
<td></td>
</tr>
<tr>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td>O</td>
<td>O</td>
</tr>
<tr>
<td>O</td>
<td>X</td>
</tr>
<tr>
<td>X</td>
<td>O</td>
</tr>
<tr>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

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

### Table 2. Statistics of validated vocabulary

<table>
<thead>
<tr>
<th>vocabulary status</th>
<th>voting</th>
<th>validating</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>accept</td>
<td>reject</td>
</tr>
<tr>
<td>SA</td>
<td>38</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>26.03%</td>
<td>100%</td>
</tr>
<tr>
<td>WA</td>
<td>108</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>73.97%</td>
<td>61.11%</td>
</tr>
<tr>
<td>total</td>
<td>146</td>
<td>104</td>
</tr>
<tr>
<td></td>
<td>71.23%</td>
<td>71.23%</td>
</tr>
</tbody>
</table>

7 Conclusion

In this paper, we developed an expertise-based vocabulary suggestion system. The goals are to realize suggester’s confidences and to reduce the load of linguists to check all accepted vocabularies. If suggesters have high confidence, their vocabularies can be conceded. Unless we utilize confidence score for checking vocabularies in system, linguists will validate laboriously.

The development is success as expect. Based on linguists’ opinion, the system reduces work and time consumption. Linguists are possible to verify vocabularies referring to strong accept level and trust a suggester who has a high confidence score.

Currently, there are 452 suggesters and 9,826 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 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

5. Longdo: http://www.longdo.com
6. LEXiTRON: http://lexitron.nectec.or.th
7. Gotoknow: http://gotoknow.or.th