LEXiTRON Vocabulary Suggestion System with Recommendation and Vote Mechanism

Kanokorn Trakultaweekoon, Peerachet Porkaew, Thepchai Supnithi
Human Language Technology Laboratory
National Electronic and Computer Technology Center
Thailand Science Park
112 Paholyothi Road, Klong 1, Klongluang, Pathumthani, 12120, Thailand
{kanokorn.trakultaweekoon, peerachet.porkaew, thepchai.supnithi}@nectec.or.th

Abstract

LEXiTRON is an online dictionary developed by NECTEC. It was originally constructed from a large corpus. However, there are many vocabularies which are not included. Thus, in this paper, we present a vocabulary suggestion system to allow users to suggest lexical data to LEXiTRON. The goals of our suggestion system are to match user interests and to minimize validation load.

This collaborative system has 3 steps to add new lexical entry. First, the user suggests a vocabulary entry with its details. This step can be done by individual suggestion or by our recommendation mechanism. Second, the suggested entry will be voted by other users. If it passes the criteria of voting, finally, it will be checked by linguist who can reject or accept it.

We confirm that the suggestion system is a promising and practical framework. Recommendation mechanism can guide contributor to suggest new item that match user interests. And, the vote mechanism can reduce validation load.

1. Introduction

The general problem of developing a dictionary is time consumption problem. Setting up a team of lexicographers for adding words and lexical item is a simple option. It guarantees that the data is correct. However, it needs a lot of lexicographer in spite of human resource lacking. Furthermore it also takes long time to define each vocabulary. In addition, we do not know whether added item will match user interests or not.

On the other hand, allowing user to submit their vocabularies and lexical data and then submit to linguists for validation will be a good option, if there are enough users in community. This collaborative framework is used in Papillion project (Mangeot and Sérasset, 2002) and Longdo (Longdo). This idea can reduce the budget and the time of improvement.

LEXiTRON is an online dictionary developed by NECTEC since 2003. The dictionary was originally constructed from a corpus which consists of frequently-used vocabularies in many topics from trusted publications. Qualitative and quantitative approaches (Palingoon, 2002) were employed to assign meanings of lexical item. Currently, the database has more than 53,000 entries of English and more than 35,000 entries of Thai. In average, there are approximately 150,000 people per month accessing to LEXiTRON.

LEXiTRON have the same problems in improving its dictionary. To deal with this problem we apply social community approach because there are a lot of users in our community.

LEXiTRON vocabulary suggestion system is an improvement of the collaborative framework. It allows contributor to suggest lexical data. Besides the correctness of data, two goals of our system are listed as follows.

- The added item should match user interests
- Validation load of linguist should be as less as possible

The rest of the paper is structured as follows. Section 2 describes the overview of system, which consists of suggestion module, vote mechanism and validation module. Section 3 provides the database design. Section 4 shows the result of system. Section 5 concludes the paper and lists up future work.

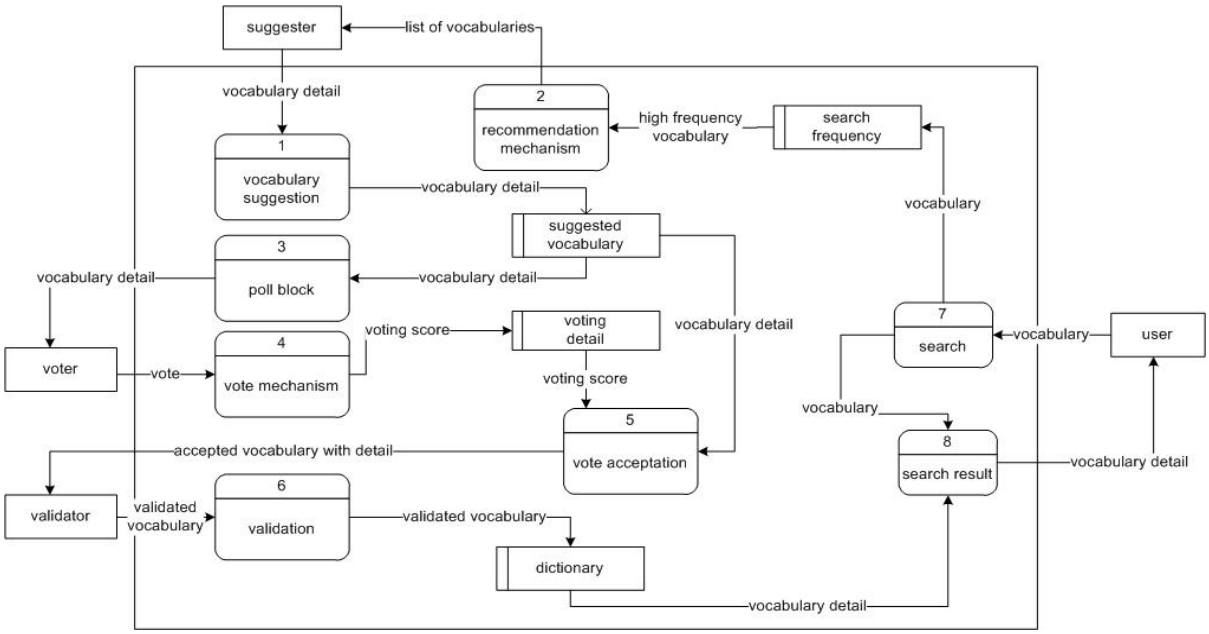

Figure 1: An overview of the system

2. System Architecture

In Figure1, our system consists of three modules; suggestion module, vote mechanism module, and validation module respectively.

Each module, represented in dashed frame, is processed by user in different roles as follows.

1. Suggester – a registered user who suggests new vocabularies. To avoid self-voting, he/she cannot vote his/her own items.
2. Voter – a registered user who votes for vocabularies suggested by other suggesters.
3. Validator – an expert of both languages who has permission to validate vocabularies accepted by voters.

The detail of each module is described in section 2.1, 2.2 and 2.3.

2.1 Suggestion Module

The objective of this module is to allow contributors to suggest new vocabularies. Contributors can submit their vocabularies with details directly.

In addition, we also provide a list of unknown vocabularies ordered by frequency. This list suggests the required vocabulary from LEXiTRON user to contributors. We named *recommendation mechanism* for this process. This mechanism helps suggester add new vocabularies that meet user interest.

2.1.1 Direct Suggestion

The process of direct suggestion is shown in Figure 2. Suggester will be asked to give some information i.e. word, part-of-speech and meaning in order to check whether this word exist in LEXiTRON or not. If the word does not exist, it will be inserted. Otherwise, it will be rejected.

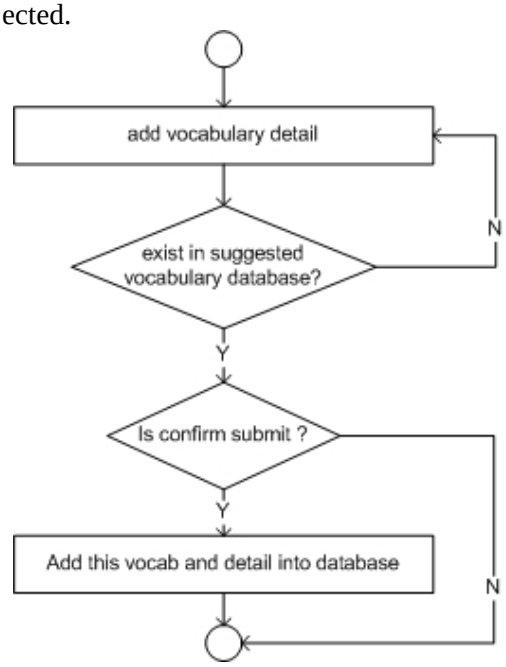

Figure 2: Flow chart of suggesting a new word

If the word does not exist in *suggested vocabulary* database, suggester has to identify category, synonym, example of usage and reference before submission. Next, the submitted word is added into *suggested vocabulary* database.

2.1.2 Suggestion using Recommendation Mechanism

In case that suggester has no vocabulary to suggest, we also provide recommendation mechanism in order to recommend some unknown alternatives. They are listed and ordered by frequency. This data is stored in *search frequency* database.

This mechanism helps us to improve our dictionary which matches user interest.

สถิติคำศัพท์ภาษาอังกฤษที่ไม่มีในเล็กซิตรอน(สถิติสะสม)	
คำศัพท์	จำนวนครั้งที่พบ
criteria	12852
exclusive	10290
evaluation	9239
exposure	8664
heard	8262
implementation	7978
wanna	7328
institute	7274
validate	5823
recognition	5574
earlier	5398
availability	5353
required	5228
innovative	4775
blog	4715
hereby	4698
regards	4212
internship	3980
recieve	3965

Figure 3: Recommendation page

Figure 4 shows the snapshot of word suggestion module. Besides the word entry, suggester has to identify some attributes i.e. *type of dictionary* (Thai->English or English->Thai), *part of speech*, *definition*, *translation or meaning*, *example of usage*, *synonyms*, *pronunciation*, *category*, *other information* and *reference*. The "*" stands for required items.

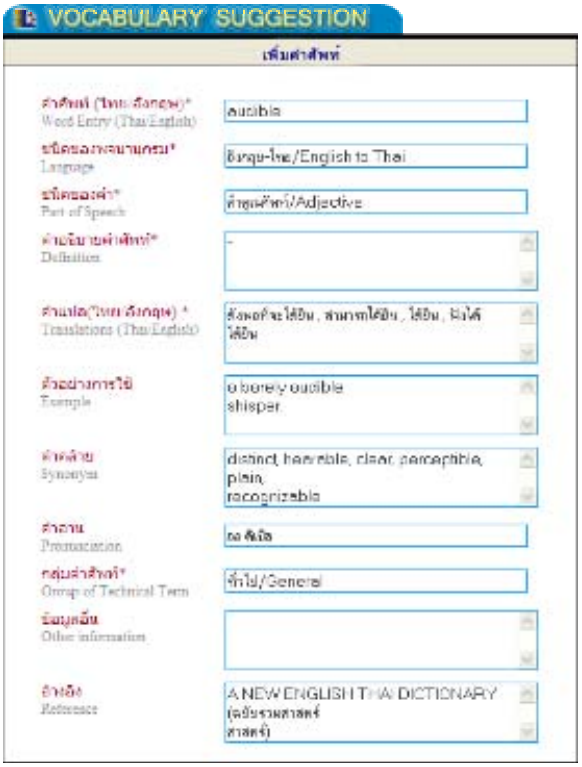

Figure 4: The snapshot of vocabulary suggestion

The different between our system and other online dictionaries is that we provide category to define the dictionary pair and domain for each word. We initially categorize the word into 20 categories i.e. *general*, *mathematics*, *science*, *engineering*, *medicine*, *biology*, *computer*, *information technology*, *material science*, *astronomy*, *economics*, *language*, *education*, *psychology*, *philosophy*, *religion*, *political*, *law*, *art*, *agricultural sciences* and *other*. This is set up to give a domain-specific meaning of each word.

2.2 Vote mechanism

After a suggester has submitted a word to any other users for voting, the voting score will be stored in *voting detail* database. This score relates to the accepted conditions. It will be validated and rechecked by validator.

Figure 4 shows the process of this mechanism. A voter is restricted to vote only once for each word. Next, we will introduce acceptance level of voter and the criteria to filter improper alternatives.

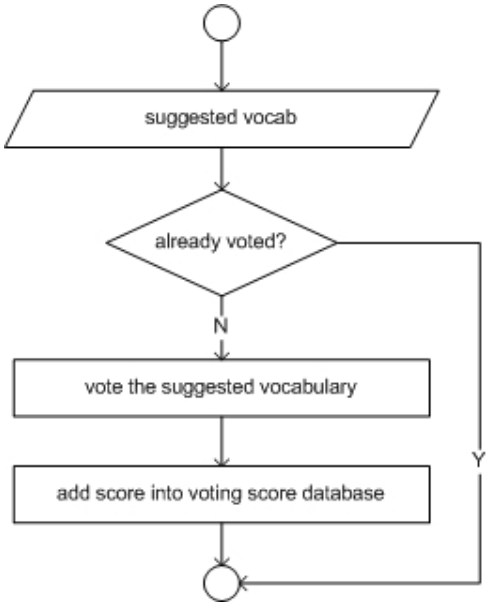

Figure 4 Flow chart of voting process

There are three levels of acceptance score.

1. Accept– if a voter votes a word to this level, it implies that the word and given detail are acceptable.
2. Reject - if a voter votes a word to this level, it implies that the given detail is incomplete and should be modified.
3. Delete – if a voter votes a word to this level, it implies that the word is not proper to be added.

The score of accept, reject and delete will be considered in order to assign status to the selected word. In our vote acceptation process, only words in status_accept and status_reject will be sent to validation module.

Figure 5 shows the pseudo code of criteria used in assigning status of each suggested entry. The constants in each condition are defined based on voter activities. It can be adjusted if the behavior of overall voter has changed.

A word can be voted by using poll block. Figure 6 shows an example of poll block displaying some information i.e. *the word, part-of-speech, meaning, example of usage, suggester's name and number of vote*. Users can see scores of each acceptance level by clicking on the number of vote. The graph of the scores will be shown in pop-up windows.

The poll block is shown in the front page after user logged in. The suggested word will be randomly displayed for vote week by week. Hence, there are only 52 words voted in a year. It is really insufficient and too slow.

```
Lets:
vote_accept = score of vote in accept level
vote_reject  = score of vote in reject level
vote_delete  = score of vote in delete level
total_vote    =vote_accept + vote_reject +
vote_delete

IF (vote_delete >=15) THEN
        vocab_status = status_delete;
ELSE IF  ((total_vote>=100) &
(vote_accept >= 0.8 * (vote accept + vote_reject +
(2 * vote_delete)) ) THEN
        vocab_status = status_accept;
ELSE IF ( (total_vote >= 100) &
(vote_reject+(2*vote_delete)>= 0.8* (vote_accept
+ vote_reject + (2 * vote_delete))) ) THEN
        vocab_status = status_reject;
ELSE
        vocab_status = status_none;
END IF
```

Figure 5: Pseudo code of our criteria

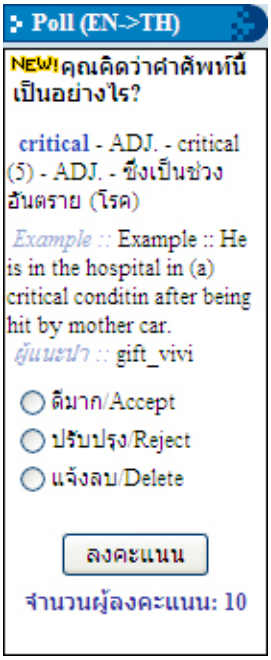

Figure 6: Example of the poll block for vote

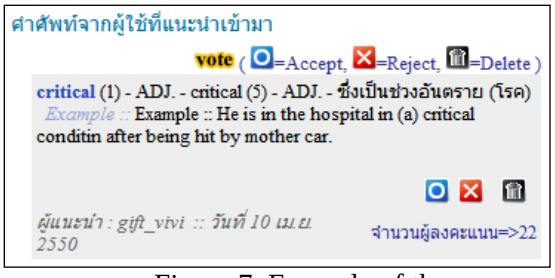

Figure 7: Example of the poll block in search page

To increase efficiency of vote mechanism, the poll block will be shown in search page as well for other choice to vote. When a user searches a word which is under inspection process, the poll block will be shown to give the optional information and to allow user vote for it.

2.3 Validation Module

In this module, validator or linguist will check and edit the words in accept and reject status. The word that passed the accept criteria will be added into database. However, depending on validator decision, not all of the word will be edited and recorded. Sometimes, improper words might pass the vote mechanism, so the validator has to delete them manually.

Figure 8 shows the flow chart of validation process. First, the validator considers whether the word is acceptable. The word will be deleted if it is not suitable or it will be added to database. Validator can edit some details as necessary. Then, the word will be added in LEXiTRON dictionary with respect the owner. (database of dictionary in Figure1)

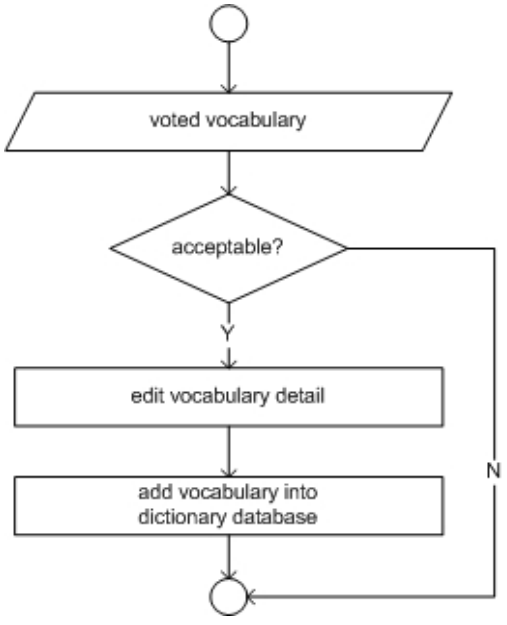

Figure 8: Flow chart of validation process.

3. Database Design

Our system consists of 4 major databases related to system overview in Figure 1.

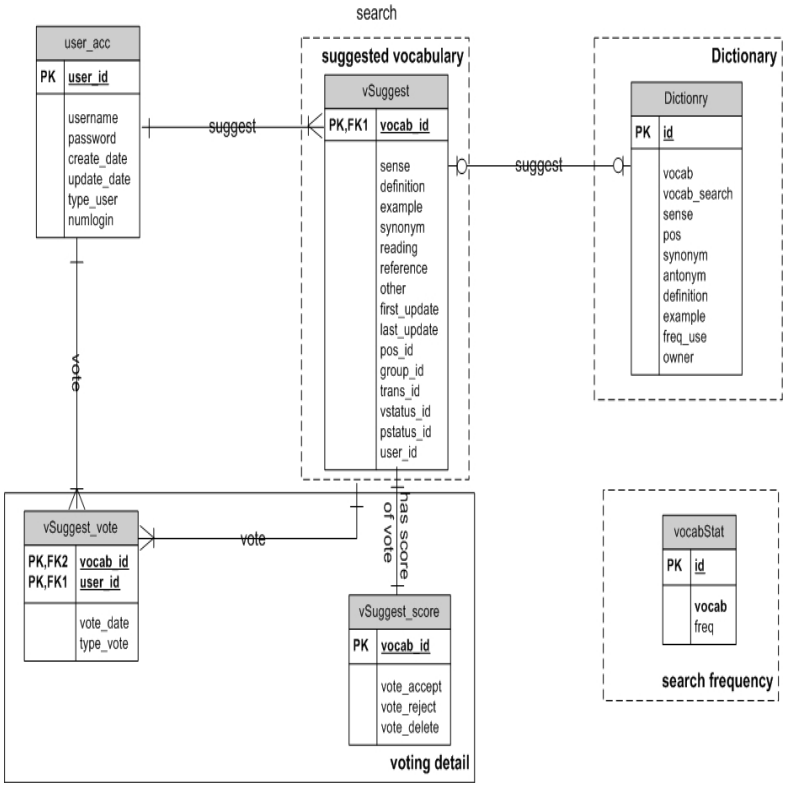

Figure 9 ER-diagram of the system

Figure 9 shows the ER-diagram of the system which consists of 6 database tables as described in Table 1.

1. user_acc	-store user information
2. vSuggest	-store detail of suggested vocabularies.
3. vSuggest_vote	-store the voting detail i.e. vocabulary, voter, date, level of voting.
4. vSuggest_score	-store accumulative score of accept ,reject and delete level.
5. dictionary	-store LEXiRON vocabulary and vocabulary added by validator.
6. vocabStat	-store search statistic of vocabularies

Table 1. he detail on table in ER-diagram

4. Result

LEXiTRON vocabulary suggestion system has been available for public since January, 2007. Currently, we obtain the following important statistics.

- There are 1,608 items suggested by 301 suggesters.
- There are 1,029 items which has been voted by 3,713 voters.
- There are approximately 4% of suggested item which was delete by vote mechanism.

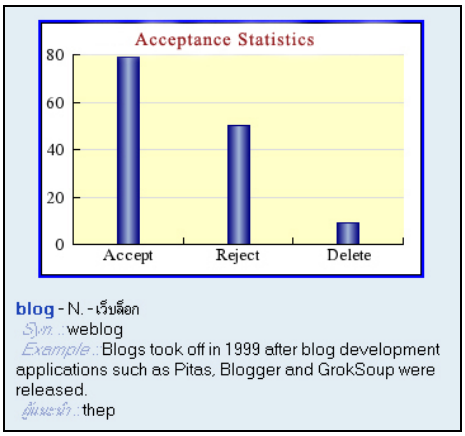

Figure 10: Example of scores in each acceptance level of word "**blog**"

Example of scores in each acceptance level of word blog is shown in figure 8. Eighty voters vote *accept*. Fifty voters vote *reject*, and ten voters vote *delete*.

- After we analyzed the deleted entries, we can group it up as follows
    - Sentence or clause – The suggested entry was not a word but it was a sentence or clause.
    - Impolite word – the suggested entry was impolite.

5. Conclusion and Discussion

We developed a vocabulary suggestion system for LEXiTRON. Our objective is to increase the size of dictionary. The system is based on collaborative framework. The goals are to improve the database in direction of user interests and to reduce the validation load of linguist or validator.

We introduce recommendation mechanism to contributor (suggester) in order to give unknown words ordered by frequency. This mechanism helps us to improve dictionary to meet user interests.

We also provide vote mechanism which allows user (in role of voter) to help validator and reduce validation load.

In fact, the efficiency of our mechanisms mainly depends on expertise, in both languages, of voters and suggesters. However, we cannot know their expertise directly. Therefore, in future work, we will develop user expertise level system to deal with inequality of their expertise.

The user expertise system will assign reliability value to each user. It will be adapted automatically to their contribution quality. For example, if his/her suggested word is accepted by voter or validator, his/her reliability value will be increased. In addition, vocabulary suggested by more reliable user should be easier to be accepted. And, this reliability value also can be a weight to score in vote mechanism.

6. References

Mathieu Mangeot and Gilles Sérasset. 2002. 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*,pages 9-15, Taipei, Taiwan, 31 August.

Pornpimon Palingoon, Pornchan Chantanapraiwan ,Supranee Theerawattanasuk, Thatsanee Charaoenporn, Virach Sornlertlumvanich. 2002. Qualitative and Quantitative Approaches in Bilingual Corpus-Based Dictionary, *Proc. Of conference of SNLP-Oriental COCOSDA 2002*, pages 152-158, Prachuapkirikhan, Thailand, 9-11 May.

Longdo: http://www.longdo.com

LEXiTRON: http://lexitron.nectec.or.th