International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:5, No:8, 2011

Lexical Database for Multiple Languages: Multilingual Word Semantic Network

K. K. Yong, R. Mahmud, and C. S. Woo

Abstract**—**Data mining and knowledge engineering have become a tough task due to the availability of large amount of data in the web nowadays. Validity and reliability of data also become a main debate in knowledge acquisition. Besides, acquiring knowledge from different languages has become another concern. There are many language translators and corpora developed but the function of these translators and corpora are usually limited to certain languages and domains. Furthermore, search results from engines with traditional 'keyword' approach are no longer satisfying. More intelligent knowledge engineering agents are needed. To address to these problems, a system known as Multilingual Word Semantic Network is proposed. This system adapted semantic network to organize words according to concepts and relations. The system also uses open source as the development philosophy to enable the native language speakers and experts to contribute their knowledge to the system. The contributed words are then defined and linked using lexical and semantic relations. Thus, related words and derivatives can be identified and linked. From the outcome of the system implementation, it contributes to the development of semantic web and knowledge engineering.

Keywords**—**multilingual, semantic network, intelligent knowledge engineering

I. Introduction

THE idea of building the Multilingual Word Semantic Network (MWSN) is inspired by the developments of semantic web and open source approach. Knowledge in the web is over flooded. Computer Scientists are coming up with many different ideas in order to represent knowledge in a better way. One of the powerful Artificial Intelligence knowledge representation techniques is semantic network. This has become the core design of Multilingual Word Semantic Network [1].

For a multilingual system, it is impossible to rely on a few contributors in database building. Knowledge sharing has become a new trend in the Internet world [2]. Thus, open source approach has become another key point in Multilingual Word Semantic Network. This system uses an open source approach as the development philosophy in database building.

Furthermore, the aim of this system is to break the boundaries in languages. Knowledge is currently represented in many different kinds of languages but pursuing knowledge should not be constrained by boundaries of languages [3]. The need to understand the relationships between words in all languages is getting more obvious due to globalization [4]. By supporting multilingual words in the database, this system will represent knowledge of different languages in a better way with the power of semantic network and open source approach.

K. K. Yong was with Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, 50603 Malaysia (e-mail: kokkhuen.yong@gmail.com).

R. Mahmud is with Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, 50603 Malaysia (e-mail: rmana@um.edu.my).

C. S. Woo is with Department of Artificial Intelligence, Faculty of Computer Science and Information Technology, University of Malaya, Kuala Lumpur, 50603 Malaysia (e-mail: cswoo@um.edu.my).

II. Related Work

WordNet is a large lexical database for English words. Nouns, verbs, adjectives and adverbs in WordNet are grouped into sets of cognitive synonyms, which are known as synsets, each expressing a distinct concept. Synsets are interlinked according to conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can then be further explored [5].

One of the strengths of WordNet is that it interlinks specific senses of words. Thus, words that are found in close proximity to one another in the network are semantically disambiguated. Besides that, WordNet labels the semantic relations among words. It works better than an ordinary thesaurus because a thesaurus usually only groups words according to meaning similarity [6].

However, WordNet only caters for words in one language, which is English. There are a few derivatives of WordNet, like EuroWordNet and GermaNet, which cater for a few European languages, but still a universal system that supports all languages is not available yet [7].

BabelNet integrates lexicographic and encyclopedic knowledge from WordNet and Wikipedia [8]. This has inspired the proposed system to utilize the power of open source in database building. By doing so, it encourages the participation and contribution of native language speakers and experts from all around the world. To build such systems, language expertise is the key to produce valid and reliable data [9].

Hence, based on the similar concept of WordNet, Multilingual Word Semantic Network is proposed to serve as a multilingual lexical database using semantic network and open source approach.

III. Background

A. Semantic Network

Semantic Network is made up of nodes and arcs. It is a kind of knowledge representation technique. Nodes represent the concepts while arcs define the relations among the nodes in the semantic network [10].

In 1956, Richard H. Richens introduced semantic network during his work in machine translation. Since then, this

approach has become an important knowledge representation technique in Artificial Intelligence, especially for Natural Language Processing and expert system development.

B. Lexical Relations

A lexical relation is defined as a pattern of association that exists between lexical units, which is culturally recognized, in a particular language. The examples of lexical relations of words are synonymy, antonymy, hypernymy, hyponymy and homonymy [11].

Lexical relations of words are important aspects in developing Natural Language Processing applications like corpus, dictionaries and thesauruses. Lexical relations offer a systematic way in representing knowledge.

C. Open Source Approach

An open source approach is a kind of practice in production or development. This approach encourages public to involve or access to the source material of an end product. This approach can be considered as a development philosophy as well as a production pragmatic methodology [12].

IV. System Design

A. Mechanism

In order to get the mechanism of the Multilingual Word Semantic Network running, English – the international language is used as the main language in the system. This means that the conducting language as well as words used in the proposed system, like the navigation links, popped-up messages, text on the search button, etc. will be in English. However, users can search words that are in other languages (users' input) and the results (system output) will be represented in multiple languages as long as there are entries in the database. The proposed system is tested and linked three different languages, namely English, Malay and Chinese Languages used by Malaysian community.

The design of the system mechanism will be based on the concepts proposed and discussed previously: using semantic network as the knowledge representation technique, open source approach in database building and multilingual words are supported in the system to break the language barrier.

Treating every word as a node, the relations among words will be the arcs. The relations that can be defined in the system are derivatives, translations, lexical relations and semantic relations.

For every word that is added into the system, the annotation of word will be done. These include its language, part of speech, pronunciation, and meaning. This helps users to understand the words during the node expansion process. Figure 1 is included in order to present visualization on how words are represented in the Multilingual Word Semantic Network:

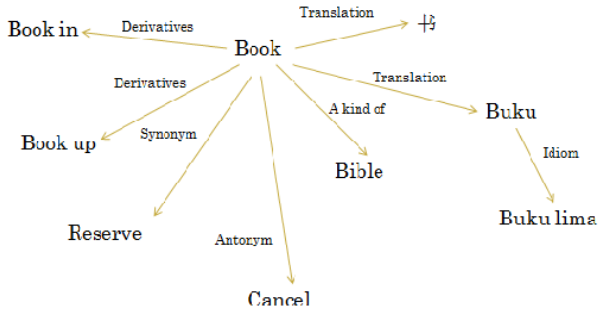

Fig. 1 How Words Linked in Multilingual Word Semantic Network

There are three modules in Multilingual Word Semantic Network, which are Open Module, Registered Module and Administrative Module.

B. Open Module

Web pages from Open Module can be accessed by any web users. The related pages are the 'Home' page and the 'About' page.

In 'Home' page, the main mechanism of word searching is done. It allows users to enter a word of a certain language and then perform node expansion from the search result. Figure 2 shows the 'Home' page of the system:

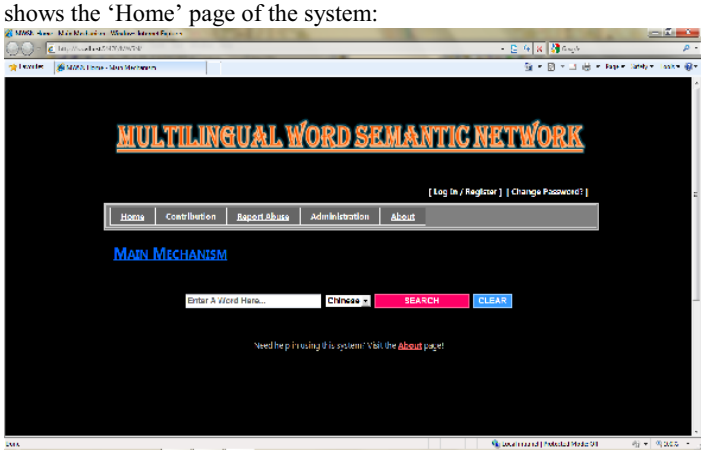

Fig. 2 'Home' page

The 'About' page is a page where explains the mechanism of the system. Functions of each section of the system were written in short instruction forms to help users to adapt to the system in a short time.

C. Registered Module

In Registered Module, only registered users are allowed to access the related pages. Basically, they are pages from the 'Contribution' tab and 'Report Abuse' tab in the hover menu.

From the 'Contribution' tab, there are pages like 'Nodes: Words' page, 'Arcs: Derivatives' page, 'Arcs: Translations' page, 'Arcs: Lexical Relations' page and 'Arcs: Semantic Relations' page. 'Nodes: Words' page allows registered users to add new words into the database. This action must be done before defining any other relations among words if the

required word is not added in the database yet. For 'Arcs: Derivatives' page, it allows registered users to define the relations of derivatives of words. For example, the word 'book up' is a derivative of the word 'book'. New relations of derivatives can be added from 'Arcs: Derivatives' page. To deal with the translations of words of different languages, here comes the 'Arcs: Translations' page. For example, the word 'book' has a Malay translation 'buku'. This relation can be added through 'Arcs: Translations' page. To define lexical relations like synonymy, antonymy and so on, it can be done in the 'Arcs: Lexical Relations' page. For example, the word 'reserve' is the synonym of the word 'book'. This relation can be added through 'Arcs: Lexical Relations' page. In order to define other relations, there is a page known as 'Arcs: Semantic Relations'. This is a page for any other semantic relation that cannot be fit into pages that previously introduced. For example, the word 'elephant' and the word 'trunk', the semantic relation is 'has', which means an elephant has a trunk. This page provides the flexibility for users.

Another page that is also under the Registered Module is the 'Report Abuse' page. This is where registered users can report faulty data or abusive user activities to the administrators.

D.Administrative Module

This module is specially designed for administrators to pre-set data, to make modification on existing data, to perform user management as well as to read the report submitted by registered users. The related pages of this module are pages from the 'Administration' tab. They are only accessible by administrators (users with administrative account).

'Reports' page from the 'Administration' tab is a page where administrators view and modify the reports submitted by registered users. Besides that, for user management, there is a page known as 'Users: Roles' page. 'Users: Roles' page is a page where administrators can assign a user to the role of an administrator or unassigned an administrator from the role of an administrator. Another page for user management is the 'Users: Deletion' page. 'Users: Deletion' is a page where administrators can delete a user account. This is very useful when dealing with abusive users.

As mentioned, administrators can pre-set data from the system. The 'Presets: Languages' page is a page where administrators can pre-set the languages that can be selected by the users during the addition of words into the system. For example, when there is a time where this system wants to support Spanish word but it was not previously, then administrators can pre-set the data 'Spanish' in this page, and thus when user wants to add in Spanish word, the option of 'Spanish' in the language field can be selected. Another page that allows administrators to pre-set data is the 'Presets: Part of Speech' page. 'Presets: Part of Speech' page is a page that deals with the presets of part of speech in the system. Next, there is a page that allows the presets of the lexical relations in the system. It is known as 'Presets: Lexical Relations'. It has the similar function to the 'Presets' pages introduced previously, except this page is dealing with lexical relations.

To modify or delete existing data, administrators need to access to 'Modification' pages. These pages are also from the 'Administration' tab. 'Nodes: Words (Modification)' page is one of the 'Modification' pages that allows administrators to modify or delete data that added from 'Nodes: Word' page. A page of the similar function is 'Arcs: Derivatives (Modification)' page. It allows administrators to modify or delete data entered from 'Arcs: Derivatives' page. 'Arcs: Translations (Modification)' page allows administrators to edit or delete translations of words that were defined in 'Arcs: Translations' page. To modify or delete data that entered from 'Arcs: Lexical Relations' page, there is a page known as 'Arcs: Lexical Relations (Modification)' page. Similar to it, 'Arcs: Semantic Relations (Modification)' page allows administrators to edit or delete data entered from 'Arcs: Semantic Relations' page.

Besides these, this system also enables administrators to modify pre-set data. For example, there is a page known as 'Presets: Languages (Modification)' page which allows administrators to edit or delete data entered from 'Presets: Languages' page. In 'Presets: Part of Speech (Modification)' page, administrators can edit or delete the data entered from the 'Presets: Part of Speech' page. Next, there is a page known as 'Presets: Lexical Relations (Modification)' page. It is a page where administrators can modify or delete data entered in 'Presets: Lexical Relations' page.

E. Additional Web Pages

In addition, besides the 3 main modules mentioned above, there are also 3 important pages that deals with user account. They are the 'Register' page, 'Log In' page and the 'Change Password' page.

'Register' page allows an unregistered user to register a new account in order to contribute to the system. For users that already have their own account, they need to sign in to access to pages from the 'Registered Module'. Administrators will also need to sign in to access to pages from 'Administrative Module'. The sign in process can be done in 'Log In' page. The last but not least, it is the 'Change Password' page which allows registered users to change their password for security purposes.

V. System Implementation

A. Search and Node Expansion

Search and node expansion is the core mechanism of the system. In order to perform a search, users need to enter a word and select the language of the word. After clicking the 'Search' button, the search result will be displayed in a list box as shown in Figure 3:

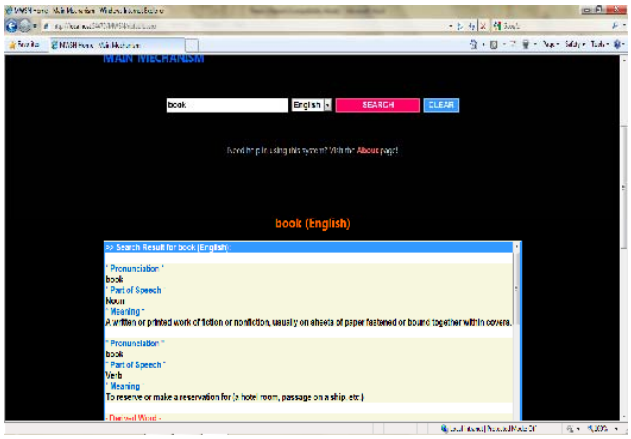

Fig. 3 Search for a Word

After the search result is displayed, there will be the node expansion options below the list box to enable users to perform node expansion. The node expansion options are 'Expand by Selection', 'Expand by Arc' and 'Expand by Random' as shown in Figure 4:

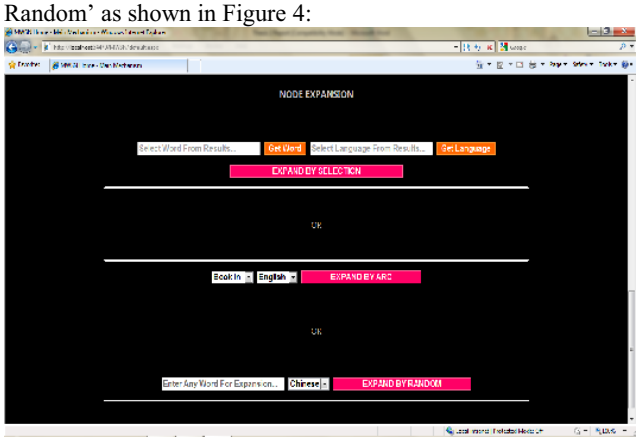

Fig. 4 Node Expansion Options

'Expand by Selection' is where users can select a word from the list box and click on the button 'Get Word', the selected word will be added to the textbox. Similar action goes for the 'Get Language' button. This feature is useful when user is dealing with non-Unicode characters or symbols.

'Expand by Arc' is a node expansion option where the system suggests nodes for users to expand. The nodes are basically the next degree of nodes to be expanded from the search result. User will have to choose from the dropdown list to perform the node expansion by arcs.

'Expand by Random' is an option which provides flexibility for users. It is similar to the search mechanism. User can enter any node into the textbox to perform the node expansion.

Finally, the expansion result will be added into the list box without clearing the initial search result.

B. New Data Addition

Registered users can add new words as well as define new relations among words in this system. The operation of adding new data into the database is rather simple.

User will have to enter the related information into the relevant textboxes, choose the related information from the dropdown lists and then click on the 'Insert' link in order to add new data into the database. The field 'Added by' will be automatically added into the system according to different user account. Figure 5 shows an example of new data addition page:

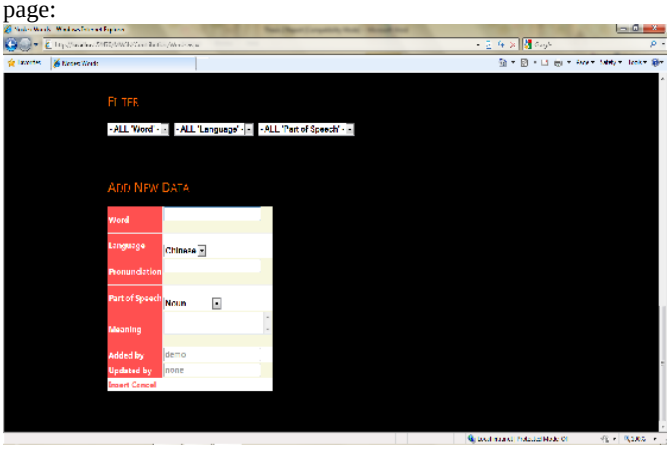

Fig. 5 New Data Addition

Similar process applies to all pages that enable users to add new data. The main point here is that users must be registered users in order to perform such operation. This is to keep track on the user activity as well as the behaviors of users in the system.

C. Existing Data Modification

In order to modify or delete existing data, an administrative account is required. The modification operation and the deletion operation are fairly straightforward. To edit data, click on the 'Edit' link and the rows of data will then be able to be edited. To delete a row of data, click on the 'Delete' link and a confirmation message will be popped-up. After clicking 'Ok', the whole row of data is then deleted. Figure 6 shows an example of existing data modification page:

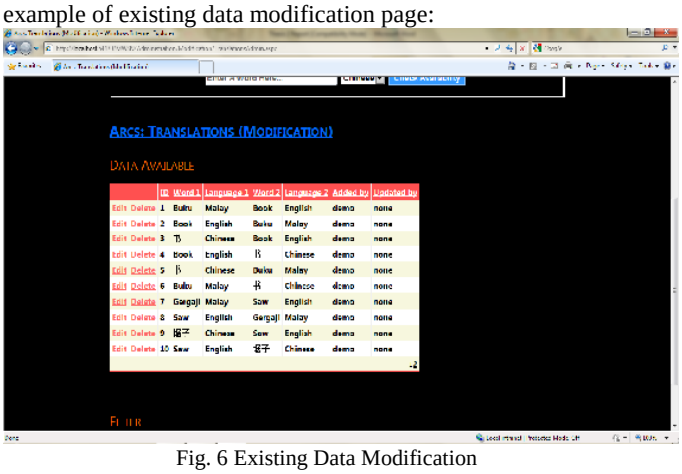

Fig. 6 Existing Data Modification

D. User Management

First of all, it is the function of assigning or not assigning a user for the role of administrator. In 'Users: Roles' page, select

a username and then click on the 'Assign User as Administrator' button will grant the selected user administrative rights.

Another function of user management provided is user deletion. To delete an abusive user, choose a username and then click on the 'Delete User' button in 'Users: Deletion' page.

E. Additional Features

Additional features were added in to the system to enhance the usability of Multilingual Word Semantic Network.

First, it is the 'Word Availability Checker'. It placed on the top of the pages that dealing with data insertion or modification. This is to ease the users to check on the word they wanted to insert into the database whether the word is already existed in the database or not. For example, when a user wants to define a relation of synonymy where the word 'hit' and 'beat' are synonyms, the user will have to add both words into the database first before defining their relations. There are too many of words in the database, hence the user can user the 'Word Availability Checker' to check whether these words existed in the database or not.

Besides this, Multilingual Word Semantic Network also provides filters for pages that deal with Grid View. It is to ease the process of finding data in Grid View as most of the data are presented to users in a Grid View form. When filters are applied on certain fields, only the selected data will be shown.

Next, there is a word suggestion function added to certain textboxes in the system. The suggested words are words that existed in the database. This will help users to reduce their time when trying to search a word.

For the dropdown lists in the system, a list search function is also added. For example, when a user wants to select a particular language from a dropdown list, it will be a tough task if the entries of languages are of a large amount of data in the database. So, in this case, user can type in the initial letter or letters of the word like 'e' or 'en' for 'English', and the selected item of the dropdown list will automatically scrolled to the word 'English'.

The last but not least, to handle users that forgot their password, a SMTP server is programmed. In the 'Log In' page, users that forgot their password will just need to enter their username in the 'Forgot Password' section. After that, click on the 'Notify the Administrator' button and a temporary password will be sent to the users' mailbox. After receiving the email which contains the temporary password, users that forgot their password will be able to log in to the system by using the temporary password and then change their password later in the 'Change Password' page.

VI. Experiment Results and Discussion

For experiments, tests had been done in areas like data duplication, data redundancy, unauthorized access and web browsers. Data duplication is checked during addition or modification of data. Same data must be avoided as it could affect the search result as well as the mechanism of the system. Data redundancy is checked to prevent errors during data deletion. Data in used should not be deleted as it will result in null reference. As Multilingual Word Semantic Network has different modules for web users, registered users and administrators, an appropriate handler is needed to deal with unauthorized access. Furthermore, the compatibility of the system with different web browsers is also tested. This system works well in all the popular web browsers. It works best in Internet Explorer 8, Mozilla Firefox 3.6 and Opera 10.

Multilingual Word Semantic Network uses semantic network as the knowledge representation technique for the input words. This approach organizes and presents data in a systematic way according to concepts and relations.

Besides that, the open source approach is applied as a development philosophy for the development of the database. It is opened for public to contribute words in building the word database as well as defining the relations among words. This is because the numbers of words and their respective semantic relations are too many to be defined. Therefore, web users, especially native language speakers or writers are the best labor power in building the database. By doing so, the system will also be able to be updated from time to time.

Finally, by supporting multiple languages, Multilingual Semantic Network offers a way to engineer knowledge of different languages. Hence, language barrier during knowledge acquisition can be eliminated through the implementation of this system.

VII. Conclusion

Large amount of data available in the web has made data mining and knowledge engineering a tough task. Next, the function of existing language translators and corpora are only limited to certain languages and domains, and search engines with traditional 'keyword' approach producing only limited results. Both of them are not efficient for knowledge acquisition.

The purpose of this project is to develop a multilingual lexical database. The aims are to tackle the problems during knowledge engineering of large amount of data in the web using semantic network as the knowledge representation technique, and to eliminate the language barrier during knowledge acquisition by supporting multiple languages and using the open source approach. At the same time, it suggests a more intelligent way to interlink words, which can then be further applied in search engines, knowledge engineering agents and the development of semantic web.

Multilingual Word Semantic Network is an innovation of WordNet. A few modifications and new ideas are proposed to overcome the weaknesses in WordNet and its relevant systems. Instead of catering only for one or a few languages, this system supports all languages. Besides that, it uses open source approach in database building to encourage knowledge sharing among native language speakers and non-native language speakers, and also among experts and ordinary users. This will also ensure that the knowledge in the system will always be up

to date.

As conclusion, by supporting multiple languages and the adaption of semantic network and open source approach, Multilingual Word Semantic Network will serve as a better lexical database.

ACKNOWLEDGMENT

This work is partially supported by research grants "Projek Khas Pembantu Penyelidik kepada Pengetua Kolej Pertama R0029/2010A" and "The University of Malaya Research Grant, ICT Cluster RG031-09ICT".

REFERENCES

[1] M. Steyvers and JB Tenenbaum. 2005. The Large-ScaleStructure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. Cognitive Science, 29: 41–78.
[2] Huysman, M. and Wulf, V. (2006) IT to support knowledge sharing in communities, towards a social capital analysis. Journal of Information Technology, 21, 40-51.
[3] C. Havasi, R. Speer, and J. Alonso. ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge. In Recent Advances in Natural Language Processing, Borovets, Bulgaria, September 2007.
[4] Darren Cook. 2008. MLSN: A multi-lingual semantic network. In 14th Annual Meeting of the Association.
[5] WordNet, "http://wordnet.princeton.edu/", last accessed in June 2011.
[6] Miller, G. A. 1995. WORDNET: A Lexical Database for English. Communications of ACM(11): 39-41.
[7] B. Hamp and H. Feldweg. Germanet - a lexical-semantic net for German. In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Re-sources for NLP Applications, Madrid., 1997.
[8] R. Navigli, S.P. Ponzetto, BabelNet: Building a very large multilingual semantic network, in: Proceedings of the 48th Annual Meeting of the Associationfor Computational Linguistics, Uppsala, Sweden, 11–16 July 2010, pp. 216–225.
[9] Hippel, E. v., How Open Source Software Works: "Free" User-to-User Assistance., Research Policy, 32, 923-943., 2002.
[10] Sowa, J. F. (Eds.). (1992). Principles of Semantic Networks. San Mateo, CA: Morgan Kaufmann Publishers.
[11] D. Hindle and M. Rooth. Structural ambiguity and lexical relations. Computational Linguistics,19(1):103–120, 1993.
[12] Open Source, “http://en.wikipedia.org/wiki/Open_source", last accessed in June 2011.