An Extracted Database Content from WordNet for Natural Language Processing and Word Games

Josephine E. Petralba
College of Information, Computer and Communications Technology
University of San Jose-Recoletos
Cebu, Philippines
josephinepetralba@apps.usjr.edu.ph

Abstract — **WordNet which is available online and in desktop applications, is an English dictionary where the synonym sets of group of words are linked by means of semantic relations such as hyponymy, meronymy and entailment, among others. The main objective of this paper is to provide the Natural Language Processing (NLP) researchers and Word Game developers with a database such that WordNet content are accessed using simple Structured Query Language (SQL) queries. A distribution copy of Wordnet 3.0 database was downloaded, and loaded into a mySQL database. It was then migrated to Oracle where the database processing to accomplish the objectives of this project was performed. There were 7 tables, 32 materialized views and 4 stored functions constructed. It is at the WordNet dictionary displays that an NLP researcher will initially investigate what Wordnet content he/she needs. Most of the objects were created with reference to the displays. The aim was to come-up with simple SQLs such that the output of an SQL is similar to what is displayed online. Queries to extract content for some Word Games such as Hangaroo™ and Batang Henyo™ (Genius Child) exemplified the use of this project for Word Games. For Oracle users, distribution copies were made available in a collection of SQL scripts. Non-Oracle users were provided with Excel spreadsheets, Comma Separated Values (CSV) and eXtended Markup Language (XML) files that they can import or load.**

Keywords- WordNet; database; Word game; download; excel

I. INTRODUCTION

A challenging aspect in natural language processing applications and word game software development is how to obtain dictionary content. Some would try to go through the task of creating content from scratch which would need bunch of researchers and content writers. A much better alternative is to extract from existing dictionaries. A good source of content is WordNet, which is a huge reservoir of semantic knowledge. It is an English dictionary where each semantic component is cross-referenced by "knotty" indices. It has the most complete collection of meaningful interrelated semantic data. WordNet's structure makes it a useful tool for computational linguistics and natural language processing. It is free, reliable and has the most correct collection, considering that it is well-reviewed by its proponents from Princeton University. It can be observed that the content in some Word Games is limited. For example, after several plays with Hangaroo, the player may notice that some hidden words are repeated. WordNet will be a rich resource for providing vast collection of hidden words for Word Games such as Hangaroo.

II. RELATED WORK

Although dictionaries contain keenly elaborated list of words with information about them, the efficient organization of that information has not been developed or used to greatest advantage in such a way as to make the information available for computer applications. There is a need to make natural language processing systems having ability of processing English word and the need to examine ways of facilitating the lexicon construction undertaking (Chodorow, Byrd and Heidorn 1985).

(Vickrey, Bronzan, Choi, Kumar, Turner-Maier, Wang and Koller 2008) acknowledges that acquiring data is of considerable important obstacle for many NLP tasks. The proponents consider the idea of creating an online game for obtaining data by collecting semantic relationships between words, such as hypernym/hyponym relations. The primary goal of the study is to produce a large amount of clean, useful data. Keeping the game fun and making sure the collected data is not too noisy were the two significant difficulties identified in the study.

There were several Wordnet projects on the construction of SQL database from Wordnet. One of these was Wordnet SQL Builder. It was distributed as a Java utility (Bou 2009). It generates the Wordnet 3.0 core database in MySQL format after selecting and downloading the modules from the download area, decompressing the zip files and executing restore-[mysql|posgresql][bat|sh].

The performance of any Natural Language Processing system is limited by the large size and complexity of collection of lexical and semantic information (Fellbaum 2010). Many natural language processing tasks, such as information extraction, sentiment analysis and word sense disambiguation makes use of WordNet database. (Pourvali & Abadeh 2012) used WordNet as one of the references for word disambiguation. A Wordnet graph, which depicts relations between words, such as hypernyms, is created. (Kubis 2012) considers the difficulty of querying WordNet databases. It suggested an approach with regard to WordNet-like lexical databases by presenting a query language, WQuery. It works with contents stored in XML files and operates on platforms that provide Java Runtime Environment.

978-1-4799-5330-1/14/$31.00 ©2014 IEEE

III. PROJECT OBJECTIVES

A. General Objectives

The main objective of this research is to provide the NLP researchers and Word Game developers with a database such that WordNet content can be accessed using SQL.

B. Specific Objectives

- To present the richness of WordNet Content for NLP and Word Games.
- To mimic through SQL, the WordNet online dictionary displays. Sample SQL queries are constructed for the different displays of the WordNet online dictionary.
- To construct database objects for Word Game applications.
- To discuss how the project database can be utilized for NLP and Word Games.
- To construct distribution copies of the project database. This will include SQL, Excel, XML, and CSV files.
- To make recommendations for related and further studies.

IV. PROJECT METHODOLOGY

The project started by downloading WordNet database content in mySQL. The result of the loading was the creation of WORDNET31_SNAPSHOT database. Oracle database was used in the preparation for the distribution copies. The mySQL objects were migrated into an Oracle schema named WORDNET3_ORA. Materialized views for the semantic relations were created based from how they are displayed online. Stored functions were created based from several online word games played. WORDNET3_FOR_GAMES_AND_NLP is the schema intended for distribution. In this schema the database structure is simplified. Building a distribution copy for Oracle users involved the collection of the CREATE and INSERT scripts. The size of the zipped distribution is 27.5 MB. It will take at most 20 minutes to load into Oracle. Excel and XML files were also generated from tables and materialized views, for loading into non-Oracle databases.

SQL file named demo.sql was written for mimicking the WordNet online dictionary displays. Another script named demoWordGames.sql exemplified the queries for some Word Games. The sample queries from these two files are simple and do not involve complex joins and subqueries.

V. QUERYING DICTIONARY CONTENTS FROM DATABASE

A. Querying synsets for a given word

The SEARCH_WORDNET table has 206353 rows. It contains the basic dictionary elements of a word. Below is a WordNet online display of the word "java"

Figure 1. WordNet online display on lemma= 'java'

Corresponding SQL query and SQL output:

select TAGCOUNT, SYNSETID, LEXDOMAINNAME, LEXDOMAINID, POS, SHOWALL, DEFINITION, SAMPLESET from SEARCH_WORDNET where lemma= 'java' order by INSTR('n,v,a,s,r,' , pos||',') ,sensenum;

TABLE 1. SQL OUTPUT ON LEMMA ='java'

TAGCOUNT	LEXDOMAINNAME	LEXDOMAINID	POS	SHOWALL	DEFINITION	SAMPLESET
2	noun.location	15	n	Java#1(java%1:15:00::)	an island in Indonesia to the south of Borneo; one of the world's most densely populated regions	
1	noun.food	13	n	coffee#1(coffee%1:13:00::),java#2(java%1:13:00::)	a beverage consisting of an infusion of ground coffee beans	he ordered a cup of coffee
	noun.communication	10	n	Java#3(java%1:10:00::)	a platform-independent object-oriented programming language	

B. Materialized Views for Semantic Relations

There were 28 materialized views constructed for each of the semantic relations (or word links). Member meronym is a semantic relation in Wordnet. Below is an online display of a member meronym of "java" and its corresponding query from a materialized view.

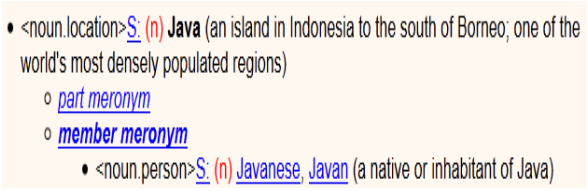

Figure 2. WordNet's member meronym for "java"

Corresponding SQL query and SQL output:

select A_LEXDOMAINNAME, A_POS, A_DEFINITION, A_SAMPLESET, C_HIDEALL,C_LEXDOMAINNAME, C_POS, C_DEFINITION, C_SAMPLESET from MV_MEMBER_MERONYM where a_lemma= 'java' order by INSTR('n,v,a,s,r,' , a_pos||',') , a_sensenum;

TABLE 2. MEMBER MERONYM ON LEMMA ='java'

A_LEXDOMAINNAME	A_POS	A_DEFINITION	A_SAMPLESET	C_HIDEALL	C_LEXDOMAINNAME	C_POS	C_DEFINITION	C_SAMPLESET
noun.location	n	an island in Indonesia to the south of Borneo; one of the world's most densely populated regions		Javanese, Javan	noun.person	n	a native or inhabitant of Java	

VI. WORDNET FOR NLP

Some Uses of WordNet in NLP:

A. WordNet is used as a source for disambiguation

Word disambiguation is applied to words that have more than one meaning. For example, the word "bank" in the sentences below has different meanings:

"Beautiful bank of river."

"He needs to pay his loans at the bank."

Tokenizing the first sentence and collecting the non-stop words will give: {"Beautiful", "bank" "river"}. Similarly for the second sentence: {"need", "pay", "loans", "bank"}.

Using NLP techniques that will perform intelligent matching of these tokens with the content from MV_DIRECT_HYPONYM, a disambiguation is able to identify that "bank" in the first sentence refers to "slope beside a body of water". In the second sentence it refers to "a financial institution".

B. WordNet is used to establish semantic distance between words.

Semantic distance refers to how close or distant the meanings between words. The Hypernym relation can be used as the basis for semantic distance. For example, for the inherited hypernym of "java", "beverage" is closer to "java" than to "substance". Inherited hypernymns are queried from the materialized view MV_DIRECT_HYPERNYM.

C. WordNet is used in Stemming

Stemming is the process for reducing derived words to their stem or root form. Some words are stemmed using simple rules such as the case for "girls", where the suffix "s" is just removed giving "girl". The MORPHOLOGY table contains those words whose stems are difficult to generalize as in the case for "men" which is "man". This table can be used to supplement several stemming algorithms on development.

VII. WORDNET FOR WORD GAMES

Most word games involve some of the following characteristics:

a) guessing a letter, meaning, or phrase related to a word or phrase;
b) rearranging of letters;
c) mapping a word or phrase into a category and its subcategories;
d) inference from some logic or related semantics of the hidden word or phrase; and
e) playing on different levels - usually a player will choose to play the easiest level first.

A. Hangaroo™

MV_HANGGAMES is a materialized view created for Hangaroo and Hangman. With the assumption that level one is the easiest and level five is the most difficult, we can partition MV_HANGGAMES according to a level of difficulty in the Hangaroo game.

TABLE 3. LEVEL 5 SAMPLE WORDS

LEMMA	SUBCATEGORY	CATEGORY	DEFINITION
annwfn	Welsh mythology	cognition	the other world; land of fairies
asymmetry	mathematics	attribute	a lack of symmetry
batsman	baseball	person	a ballplayer who is batting
gas phlegmon	pathology	state	a deadly form of gangrene usually caused by clostridium bacteria that produce toxins that cause tissue death; can be used as a bioweapon
honky	slang	person	offensive names for a White man
phylogeny	biology	process	the sequence of events involved in the evolutionary development of a species or taxonomic group of organisms
shimchath torah	Judaism	time	a Jewish holy day celebrated on the 22nd or 23rd of Tishri to celebrate the completion of the annual cycle of readings of the Torah
supersymmetry	physics	cognition	a theory that tries to link the four fundamental forces
wynfrith	Roman Catholic Church	person	Anglo-Saxon missionary who was sent to Frisia and Germany to spread the Christian faith; was martyred in Frisia (680-754)

TABLE 4. WORD COUNT FOR EACH LEVEL

LEVEL	Word Count
1	729
2	1013
3	902
4	1056
5	201

The word count is a proof of the usability of WordNet as a source of content at a cheaper cost. It would take time for a copy writer or a content expert to come-up with this number of words classified according to difficulty level.

B. Hangman™

In Hangman, the guesser first chooses a CATEGORY then guess the words belonging to that CATEGORY. A materialized view MV_INSTANCE_HYPONYM has word collection of countries, authors, musicians, constellations, among others. Another source is MV_HANGGAMES which has categories such as zoology, figurative, basketball, astronomy, Greek mythology, among others. SEARCH_WORDNET table has more generalized categories such as weather, shape, feeling and time.

C. Boggle™ , Scrabble™ and WordNet

The scrabble official dictionary was published by the National Scrabble Association (NSA). The set of words from the NSA is not the same with the set of words from WordNet. Boggle also uses the dictionary of the NSA. There is no point in extracting from WordNet the "official" words for Scrabble and Boggle, unless we need to innovate Scrabble and Boggle such that WordNet dictionary is the official reference. A materialized view

MV_WORDNET_SCRABBLE is created for this. Just like the standard scrabble, in this materialized view, proper noun such as “Lincoln” or “Daniel” is excluded.

D. *Batang Henyo™*

Batang Henyo (Genius Child) is a parlor game where effective deductive reasoning is the strategy to win. A hidden word and its category are revealed to the guessers and are hidden from the questioner. The questioner asks questions that can be answered with "Yes" or "No". The hierarchical nature of the content of materialized view MV_PART_MERONYM, makes it an appropriate reference for deductive analysis.

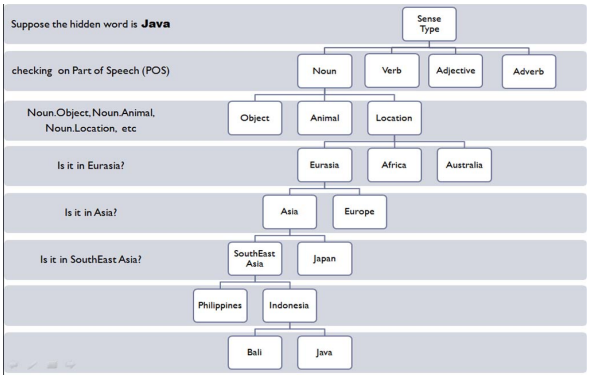

Figure 3. Logical structure on how a hidden word “java” can be deduced

The project's database can also complement with other sources of content. For example, for those words whose LEXDOMAINNAME is 'noun.place', a more efficient search can further proceed using a Goggle Map database.

VIII. Conclusion and Future Work

This paper was able to present the richness of WordNet content for NLP and Word Games as exemplified on what tables or materialized views are sourced for their needed contents. Sample queries were provided in demo.sql and demo WordGames.sql files. Distribution copies were created for those who will be using the extracted database content of this project. Oracle CREATE and INSERT scripts are generated for loading into an Oracle DB. For loading into non-Oracle DB, files in Excel, CSV and XML format were distributed.

This research will serve as a window to more related researches on Word Games, NLP or an improvement of this research itself. Here are some of the recommended related projects for future work.

1. Innovate existing word games, and design new ones. The availability of rich content from this project would create interest for the development of new word games or innovations of existing ones. For example, the current Hangaroo game can be innovated to be more educational by displaying the definition or parts of speech. The invention of new word games will draw advantages on the availability of words and their comprehensive relations with other words. One can come-up with a game that may involve word relation such as the meronyms (is a part of).
2. Undertake NLP research utilizing the extracted WordNet content in this project. The project's database can also complement with other data sources, taxonomies, dictionaries and thesaurus.
3. Develop a desktop application, web-based application or a web service interface where we can execute the demo queries in a more user friendly environment.
4. Generate CREATE and INSERT scripts for distribution for Non-Oracle databases. Currently, the set of SQL files for loading is written in Oracle SQL syntax only. It is recommended to generate for DB2, mySQL, Microsoft SQL Server and other databases.

Acknowledgment

I would like to thank the University of San Jose Recoletos administration for all the support.

References

[1] Miller, G. "WordNet-About Us, WordNet, Princeton University." (2010).
[2] Princeton University "About WordNet." WordNet. Princeton University. 2010. <http://wordnet.princeton.edu>
[3] Fellbaum, C. (Ed.). WordNet: An electronic lexical database. MIT Press. (1998). Retrieved from http://lib.freescienceengineering.org/view.php?id=474911
[4] Bou, Bernard. "Wordnet SQL builder." (2009). Retrieved from http://wnsqlbuilder.sourceforge.net/
[5] Pourvali, Mohsen, and Mohammad Saniee Abadeh. "Automated text summarization base on lexicales chain and graph using of wordnet and wikipedia knowledge base." *arXiv preprint arXiv:1203.3586* (2012). Retrieved from http://arxiv.org/ftp/arxiv/papers/1203/1203.3586.pdf
[6] Fellbaum, Christiane D. "Harmonizing wordnet and framenet." *Advances in Natural Language Processing*. Springer Berlin Heidelberg, 2010. 2-2.
[7] Chodorow, Martin S., Roy J. Byrd, and George E. Heidorn. "Extracting semantic hierarchies from a large on-line dictionary." *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 1985. doi:10.3115/981210.981247
[8] Vickrey, David, et al. "Online word games for semantic data collection." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008.
[9] Kubis, Marek. "A query language for wordnet-like lexical databases." *Intelligent Information and Database Systems*. Springer Berlin Heidelberg, 2012. 436-445.