ISSN: 2277-3754
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 1, Issue 4, April 2012

A Survey: Structure of Machine Readable Dictionary

Avanti M. Deshpande

Abstract* – MRD (Machine Readable Dictionary) is a huge lexical database of English Nouns, verbs, adjectives and adverbs which are merged into sets of cognitive synonyms (synsets), each one represents a unique concept. Synsets are joined using conceptual-semantic and lexical relations. The representation is on the level of word meanings called as lexemes. It can determine the meaning of word by its sets of synonyms. Synset has a unique index and shares its properties whereas both nouns and verbs arranged into hierarchies, representing "IS A" relationships. The relationships among the noun synsets can be interpreted as "SPECIALISATION" relations based on the concept. Hence this structure is a useful tool for computational linguistics and NLP which can be exploited as a lexical ontology.

Index Terms- **Antonym, Holynym, Hypernym, Lexicon, Natural Language Processing, Synsets, Word Sense Disambiguation.**

I. INTRODUCTION

- MRD is a machine readable dictionary which consists of lexical databases which groups English words into sets of synonyms called Synsets, which gives simple meanings of the word. It maintains various semantic relations among synonym sets.
- It maintains rich vocabulary organizational structure. Using this it is very easy to construct and expand a domain lexicon. It provides rich semantic relations of words including synonym, antonym, and so on with which words are linked together to form a network. It is a structural notion where the meaning of a concept determined using its position relative to the other words in the structure.
- It is a dictionary in an electronic form that can be loaded in a database and can be queried via application software. It may be a single language explanatory dictionary or a multi-language dictionary to support translations between two or more languages or a combination of both.
- An MRD may be a dictionary with a proprietary structure that is queried by dedicated software (for example online via internet) or it can be a dictionary that has an open structure and is available for loading in computer databases and thus can be used via various software applications.

II. NEED FOR MRD'S

Conventional dictionaries contain a lemma with various descriptions. A machine-readable dictionary may have additional capabilities and is therefore sometimes called a Smart Dictionary. An example of a smart dictionary is the Open Source Gellish English dictionary. Machine-readable versions of everyday dictionaries have been seen as a likely source of information for use in natural language processing because they contain an enormous amount of lexical and semantic knowledge. Machine-readable versions of everyday dictionaries have been seen as a likely source of information for use in NLP because they contain an enormous amount of lexical and semantic knowledge collected together over years of effort by lexicographers. Interest in machine-readable dictionaries (MRDs) as a ready-made source of knowledge has waned somewhat in recent years. Considerable research has been devoted to devising methods to extract this information from dictionaries based on the supposition that it is sufficient to form the kernel of a knowledge base that can be extended by utilizing information from other sources.

III. LITERARURE SURVEY

- **Natural language processing** (**NLP**) is a field of computer science and linguistics concerned with the interactions between computers and human (natural) languages. It began as a branch of artificial intelligence. In theory, natural language processing is a very attractive method of human–computer interaction. Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications. The goal of the Natural Language Processing (NLP) group is to design and build software that will analyze, understand, and generate languages that humans use naturally, so that eventually you will be able to address your computer as though you were addressing another person. This goal is not easy to reach. "Understanding" language means, among other things, knowing what concepts a word or phrase stands for and knowing how to link those concepts together in a meaningful way. It's ironic that natural language, the symbol system that is easiest for humans to learn and use, is hardest for a computer to master. Long after machines have proven capable of inverting large matrices with speed and grace, they still fail to master the basics of our spoken and written languages. It is a form of human-to-computer interaction where the elements of human language, be it spoken or written, are formalized so that a computer can perform value-adding tasks based on that interaction.

- **Word-sense disambiguation** (WSD) is an open problem of natural language processing, which governs the process of identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings (polysemy). The solution to this problem impacts other computer-related writing, such as discourse, improving relevance of search engines, anaphora resolution, coherence, inference *etc.*

IV. STRUCTURE OF MRD

The lexical database for English language acts as a machine-readable dictionary, created and maintained by George Miller at the Cognitive Science Laboratory at Princeton University. It is an online lexical database designed for use under program control [1].These database group English words into sets of synonyms called synsets, provides short, general definitions, and records the various semantic relations between these synonym sets. Each such synset represents a single distinct sense or concept. For example, in MRD, the synset {car, auto, automobile, machine, motorcar} represents the concept of "4-wheeled motor vehicle. There are various semantic relations among synonym sets which serve as a combination of dictionary and thesaurus for automated text analysis. MRD stores information about words belong to four parts–of–speech: nouns, verbs, adjectives and adverbs.MRD's database groups English nouns, verbs, adjectives, and adverbs into synsets that are in turn linked through semantic relations that determine word definitions and senses.

- Organized around the notion of synsets (sets of synonymous words) Groups words together based on their meanings. It interlinks specific senses of words.
- Basic semantic relations between these synsets that labels the semantic relations among words.
- Noun hierarchy partitioned into separate hierarchies with unique top hypernyms.
- Neighbouring words will be more closely related to correct sense Words that are found in close proximity to one another in the network are semantically disambiguated.

MRD is an electronic lexical database, which is arranged semantically and contains nouns, verbs, adjectives and adverbs. Words that are synonymous are grouped together in Synonym sets are referred as synsets. Each synset has an associated definition named gloss, which is a short explanation of the meaning of the concept represented by the synset. Many words in MRD are polysemous i.e. they are included in more than one synsets. For example, the word computer can be found in the synset {computer, computing machine, computing device, data processor, electronic computer, information processing system}, which has the gloss "a machine for performing calculations automatically", and in the synset {calculator, reckoner, figurer, estimator, computer} which has the gloss "an expert at calculation (or at operating calculating machines)".Synsets are connected to each other through various semantic relations. The most important relations between nouns are the relations of hyponymy and hypernymy, which are transitive relations between synsets. The hypernymy relationship between synsets A and B means that B is a kind of A. Hypernymy and hyponymy are inverse relationships, so if A is a hypernym of B, then B is a hyponym of A. For example the synset {computer, computing machine, computing device, data processor, electronic computer, information processing system} is a hypernym of the synset {home computer}. Usually each synset has only one hypernym, therefore this relation organizes WordNet into a hierarchical structure. Another pair of inverse relations that hold between nouns is the meronymy and the holonymy relations. If A is a holonym of B (or in other words B is a meronym of A), it means that B is a part of A. For example, synset {keyboard} is a meronym of the synset {computer, computing machine, computing device, data processor, electronic computer, information processing system [1].

V. ENHANCED FUCTIONALITIES

MRD evaluates concept maps designed by learners and for supporting the learner in his/her attempt to design concept maps. Sets of synonyms gives flexibility in assessing learners' answers, the ontology helps the expert to define multiple possible correct answers and finally WordNet can be used to provide meaningful feedback targeted at the learners' specific mistakes. Specifically, the synsets in MRD can provide flexibility when comparing an answer given by a learner to the correct concepts provided by the expert. Use of synsets allow us to check whether the learner's and the expert's map actually refer to the same concept in different words. For example, if a concept in the expert's map is the {central processing unit, CPU, C.P.U., central processor, processor, mainframe} synset, it makes no difference whether the learner phrases the answer as processor or CPU or C.P.U., because all these answers will be considered as correct. It gives an extensive ontology that the expert can use in order to define entire classes of correct concepts instead of one correct concept. For example in the expert's concept map the relationship "is composed of" may exist between the concept computer and a concept that the learner should fill out. Then if the learner gives as an answer any part of the computer, that answer should be considered correct, regardless of the specific part. This can be achieved through MRD's meronymy relation. Once the expert has defined the correct answer as any meronym of a certain synset, the system can evaluate the answer of the learner by searching in synsets that are meronyms of the given synset. Like this various acceptable answers can be defined by their common property and not by listing them explicitly. This seems to be well suited for concept maps, since a number of relationships that appear often in concept maps, correspond to MRD's relations. For example, the concept map relationship "is composed of" corresponds to meronymy and the "is a kind of" relationship corresponds to hypernymy. Its ontology and relationships may further be exploited so as to provide the learners with meaningful feedback that will be targeted at the specific errors they have made, instead of just generic messages such as "Wrong Answer" and "Try Again". Based on the relationship one can provide the learners with the most appropriate feedback that will hint towards the correct answer. For example, if the correct concept expected is defined as the synset {memory, computer memory, storage, computer storage, store, memory board} and the answer of the learner is RAM. Using the MRD's ontology it will be easier to find the synset {random-access memory, random access memory, random memory, RAM, read/write memory. It contains RAM which is a hyponym of the correct synset. Through this knowledge, the system guide learners towards using the general category to which RAM belongs, i.e. "memory" [4].

VI. FEATURE SELECTION USING MRD

The feature selection process using MRD is employed to discover synonymous terms based on cross-referencing. Compare the MRD's synonyms approach with statistical methods such as Chi2 and IG. IG is a feature selection technique makes use of the presence and absence of a term in a document to select its features. Chi2 measures the degree of independence between a term and a category. It selects features with high dependency on a particular category. The MRD's synonyms approach will explore the use of synonyms and word senses to derive a better set of features for category representation to achieve better categorization effectiveness [1]. In synonyms approach, feature selection is based on terms with overlapping word senses co-occurring in a category. The co-occurrence of terms with the same synset signature is used as an indicator of significant terms to represent a category. For terms co-occurring in a category, the correct sense is determined based on the synset signature cross-referencing. Cross referencing is done by checking the list of noun synsets. Proceedings of for all senses for similarity in the signatures. The senses of different terms with overlapping synsets aggregate the semantic context of a category. The original terms from the category that belongs to the similar synsets will then be identified and added as features for category representation. Let us look at all the senses for five nouns; 'corn', 'maize', 'acquisition' and 'ship'. Each sense has a signature, which is referred to as a synset containing synonyms to reflect a sense [5].

VII. APPLICATION AREAS

Machine-readable versions of everyday dictionaries have been seen as a likely source of information for use in natural language processing because they contain an enormous amount of lexical and semantic knowledge. No comprehensive evaluation of machine-readable dictionaries (MRDs) as a knowledge source has been made to date, although this is necessary to determine what, if anything can be gained from MRD research.

- **Using MRDs in Text to Speech:** Dictionary entry consists of several fields of information; naturally, each will be user-id for different applications. Among the standard fields are pronunciation, etymology, subject field notes, definition fields, synonym and antonym cross references, semantic and syntactic comments, run-on forms, conjugational class and inflectional information where relevant, and translation for the bilinguals dictionaries. Each of these fields has proven useful for different applications, such as for building semantic taxonomies and machine translation. The most directly usefully for TTS is the pronunciation field.
- **Using MRDs the Pronunciation Field:** Extracting the pronunciation field from an MRD is one of the most obvious uses of a dictionary. Nevertheless, parsing dictionaries in general can be a very complex operation and even the extraction of one field, such as pronunciation, can pose problems. Several pronunciations can be given for a headword and the choice of one must be made. Moreover, because of the rich morphology of French this has a rough ratio of eight morphologically inflected words for one base form. If pronunciation varies during inflection of nouns and adjectives, the pronunciation field reflects that variation

which makes the information difficult to extract automatically.

VIII. CONCLUSION

The basic notion of meaning used in MRD is lexical meaning, and MRD's main SYNSET relation is denoting coincidence of lexical meaning. Synset has a unique index and shares its properties whereas both nouns and verbs arranged into hierarchies, representing IS A relationships. This structure is a useful tool for computational linguistics and NLP which can be exploited as a lexical ontology. A machine-readable dictionary may have additional capabilities and is therefore sometimes called a smart dictionary. It is used for a number of different purposes in information systems, like word sense disambiguation, information retrieval, automatic text classification & automatic text summarization. Using application program the MRD will play an important role for automatic crossword puzzle generation [2].

REFERENCES

[1] Miller G. A., Beckwidth, R., Fellbaum, C., Gross, D. and Miller, K. J., "Introduction to WordNet: An On-line Lexical Database", International Journal of Lexicography, Vol 3, No.4 (Winter 1990), pp. 235-244.

[2] Aas, K and Eikvil, L. "Text Categorization: a survey", Technical Report #941, Norweigan Computing Center, 1999.

[3] Yang, Y. and Liu, X., "A re-examination of text categorization methods", in proceedings of the SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval, Berkeley, CA, 1999, pp. 42—49.

[4] Christiane Fellbaum, editor. 1998. WordNet: An Electronic Lexical Database, Language, Speech, and Communication Series. The MIT Press, Cambridge MA.

[5] George A. Miller. 1990. WordNet: An on-lie lexical database. International Journal of Lexicography, 3(4):235–312.

[6] Angles and C. Gutierrez. 2005. Querying rdf data from a graph database perspective.2nd. European Semantic Web Conference (ESWC2005), Heraklion, Greece, 3532:346{360, May.

[7] Fellbaum, C., 1998. An Electronic Lexical Database, MIT Press, Cambridge, Mass.

[8] Chua, S. and Kulathuramaiyer, N., 2004. Semantic Feature Selection Using WordNet. In Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI 2004), page(s): 166 – 172, 2004.