Acquisition of lexical units in Ontological Semantics

Amaal S.H. Al-Hashimy
Sultan Qaboos University
Computer Science Department
Oman , SQU,P.C. 123, P.O.BOX 36
amaalh@squ.edu.om

ABSTRACT

This paper is an intellectual overview to the theory of ontological semantics (OS) for natural language processing. It reviews the fundamental premises of the theory (detailed description can be found in [5]) and focuses on a crucial issue to OS which is ontological semantic lexicons, one of the static knowledge sources that OS theory depends on. Ontological semantic is an approach to developing an exhaustive and detailed linguistic theory of meaning that is sufficient for NLP (natural language processing) by computers. It is a knowledge based system that required a vast amount of information regarding the world around a specific domain of application.
This vast amount of information is encoded into the ontology and the lexicon mainly. And since the lexicon is the main concern here, the focus will be on some specific aspects that are key to the development of it, such as the acquisition of lexical units information, and the organization of the lexicon. And some modern issues like the possibility of automation of static knowledge acquisition.

Key Words: Ontological semantic theory, lexicon, lexical acquisition, automatic acquisition.

1. Introduction

Several decades of work in NLP have clearly demonstrated the importance of meaning representation of natural language as the crucial element for further high-end NLP applications (most NLP applications like MT, QA, text summarization, etc. stand to benefit from being able to use text meaning). But unfortunately the work in this field, over the decades, pertained to treatment of meaning because of the complexity of such task as providing a comprehensive meaning representation of NL is unreachable. Although the workers in this field NLP tried to overcome such obstacle by finding many theories and techniques that used in many applications, but always the outcome of such applications was very restricted and limited in its ability of handling different language aspects especially meaning.

Many researchers tried to bypass the need of true meaning treatment like using complex counts of frequencies of occurrence for strings in various contexts, shallow parsing, etc, but in most cases meaning was handled in a nonsystematic way which yield poor results. Of course most of the work done in this area is highly effective in improving the processing in this field but the problem of meaning still standing. In this case Ontological Semantic (OS) theory arose as a theory of comprehensive approach to the treatment of text meaning by computer.

The theory of ontological semantics is built as a society of micro theories covering such diverse ground as specific language phenomena, world knowledge organization, processing heuristics and issues relating to knowledge representation and implementation system architecture.

2. Ontological semantics

Ontological semantic (OS) is an approach to developing an exhaustive

and detailed linguistic theory of meaning that is sufficient for NLP (natural language processing) by computers. It is responsible of all the processes and the knowledge sources that are required in a comprehensive framework to represent natural language meaning.

The goal of ontological semantics is the extraction, representation and manipulation of meaning in natural language texts with a view toward supporting applications such as MT or question answering, information extraction, text summarization, etc. Text meaning is represented in text meaning representations (TMRs) that are derived compositionally, primarily from meanings of words and phrases in the text, where word and phrase meaning is encoded in the *ontological-semantic lexicon* (see Figure 1) . Central to this goal is the employment of the ontology, or a constructed model of the world, as a language-independent static resource, which is used to construct text meaning representation TMR of the input texts.

2.2 Main processing in OS

As any semantic theory for natural language processing, OS must account for the processes of generating and manipulating text meaning. An accepted general method of doing this is to describe the meanings of words and, separately, specify the rules for combining word meanings into meanings of sentences and, further, texts; hence the division of semantics into lexical (word) semantics and compositional (sentence) semantics. Semantics for NLP must also address issues connected with the meaning-related activities in both natural language understanding and generation by a computer.

So the meaning representation of a text is derived through:

1. establishing the lexical meanings of individual words and phrases comprising the text;
2. disambiguate these meanings
3. combining these meanings into a semantic dependency structure SDS.

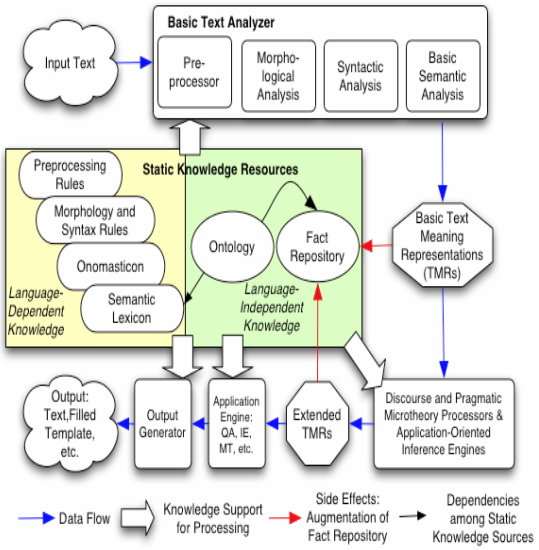

Figure 1.Overall Architecture of a Generic Application of Ontological Semantics

The SDS-building process relies on meanings of lexical units, as defined through links to the ontology and by non-propositional meaning elements; so the process is guided by the syntax-semantics interface manifested in the lexical syntactic and lexical semantic specification of lexical entries. [2]

Proceeding from the lexical, morphological and syntactic information available after the preprocessing stage(though OS is a meaning oriented approach, but it can not dispose the role of non-semantic modules like tokenizer which identifies tokens for further processing; this involves a number of auxiliary tasks, using language-specific ecological constraints) for a textual input, on the one hand, and an empty TMR template, on the other, OS starts the propositional structure of the future TMR (a computational representation of the text meaning). Words are

looked up in the lexicon and onomasticon. The parsing is a recursive process: in cases, where it is impossible to find a matching lexical entry, the restrictions on the conceptual connections are relaxed, and the process is then repeated. As a result, TMRs largely consist of instances of ontological concepts. Some of these instances are remembered (as "facts") and stored in the *fact repository*, FR, a knowledge base of remembered ontological instances. Some facts in the fact repository are referred to by proper names in texts–personal names, names of organizations, specific artifacts, etc. These proper names are stored in the *onomasticon*, the semantic zones of whose entries contain a pointer to a corresponding fact repository element. Once the TMR for the document is acquired, it can be used for a number of purposes from translation and information extraction and data mining.

Thus the main parts of the analysis process may include the following components:

1)a tokenizer that divides the input text into a series of strings and deals with all special ecological knowledge including characters, numbers, symbols, punctuation, etc.

2) a morphological analyzer, processing the inflected forms of a lexical units and establishing their meaning to be further used in text meaning representation;

3) a semantic analyzer, it is responsible of establishing propositional dependencies (using the syntactic information built into lexical item in the lexicon and semantic constraints), and also deals with the pragmatic aspects of the text: style, speaker attitude and goals, etc.

4) a module that uses a special format for the formulation of text meaning representation using the information from the previous three module.

2.3 Knowledge sources in OS

The methodology of ontological semantics consists of acquisition of the static knowledge sources (ontology, lexicon, onomasticon, and fact database) and of the procedures for producing and manipulating TMRs**.**

An implemented system of OS employs the following resources:[3]

1) The ontology, a store of general concepts in a specific domain which is language-independent, (world knowledge)
2) Lexicons of the languages, which maintain the basic words of a domain with their syntactic and semantic features.
3) Fact database and onomasticon, or depository of proper names, which contain instantiations of ontological concepts;
4) Text processing modules, most prominently a semantics text analyzer (which is intended for constructing text meaning representations from natural language texts) and semantic text generator (intended for a reverse process, constructing natural language texts on the basis of text meaning representations).

3. Lexicons in OS

Natural language processing (NLP) systems vary in their applications, and as such vary in what they require from the lexicon. The computational lexicon is the fundamental store of information about the primary component of language, i.e. words, and therefore critical for systems which aim to handle some aspect of natural language. Two key issues for the lexicon in NLP tasks are *lexical representation* and *lexical acquisition*.

The ontological semantic lexicon specifies what concept, concepts, property or properties of concepts defined in the ontology must be instantiated in the TMR to account for

the meaning of a particular lexical unit of input.

3.1 Lexical syntactic specifications

Each lexicon entry is comprised of a number of sections corresponding to the various types of lexical information.

1. General: word class, definition, example, comments, variants.
2. Syntax: syntactic dependency.
3. Semantics: lexical semantics, meaning representation.
4. Linking: case roles.

The following scheme, in a BNF-like notation, summarizes the basic lexicon structure. (see Figure 2).

```
   lexeme ::=
CATEGORY: {syn-cat}
ORTHOGRAPHY:
            VARIANTS: "variants"*
            ABBREVIATIONS: "abbs"*
PHONOLOGY: "phonology"*
MORPHOLOGY:
            IRREGULAR-FORMS: ("form"
                              {irreg-form-name})*
            PARADIGM: {paradigm-name}
            STEM-VARIANTS: ("form" {variant-name})*
ANNOTATIONS:
            DEFINITION: "definition in NL" *
            EXAMPLES: "example"*
            COMMENTS: "lexicographer comment"*
            TIME-STAMP: {lexicog-id date-of-entry}*
SYNTACTIC-FEATURES: (feature value)*
SYNTACTIC-STRUCTURE: f-structure
SEMANTIC-STRUCTURE: lex-sem-specification
```

Figure 2. Lexicon entry

The contents of the SYN-STRUC zone of a lexicon entry are an indication of how the lexeme fits into parses of sentences. In addition, this zone provides the basis of the syntax-semantics interface. Thus a brief specification of this zone is necessary to present the foundation of the semantic analysis process, which relies on the syntax-semantics interface as one of the dynamic knowledge sources used in constructing a semantic representation (i.e., the TMR) for input text.

3.2 Lexical semantic specifications

The *lexical semantic specification* found in each entry in the lexicon is the repository of low-level semantic information. The *syntax-semantics interface* links into that specification, guiding the search process by suggesting what element is a candidate for combination with what other element, and in what relation.

The base case of this specification is an indication that the word refers to a concept from the ontology, and in the process of semantic analysis, the word would result in an instantiation of that concept. In many cases that concept has further constraints on the allowable fillers for various slots or specific values filled in for literal (non-relational) slots. Some lexical semantic specifications include multiple concepts to be instantiated in a particular structure (i.e., one instantiation will be specified to be the head, and another as a filler of a particular slot). Other lexical semantic specifications might not invoke the instantiation of a concept, but just provide filler information for another concept (the adjective *blue*, for example) or relate two other concepts to be instantiated by other words. Interwoven with these semantic specifications is the syntax-semantics interface component. Particular slots in the specification may have a reference variable as the filler; the variable is bound to a headed syntactic structure during processing, and the instantiated concepts that result from the semantic processing of that syntactic structure are inserted into the indicated slot's value.

For example (figure 3) the verb "said" could be exist in two different syntactic constructs in "Hasan *said* a word", and "Hasan *said* that he will go to school". The sem-struc represents the meaning for both as the ontological concept INFORM, with its agent slot

filled with the meaning of the subject of both constructions, its theme constrained to the meaning of the object in the first construction and of the complement, in the second.

```
inform
  definition      "the event of asserting
                  something to provide
                  information to another
                  person or set of persons"
  is-a            assertive-act
  agent           human
  theme           event
  instrument      communication-device
  beneficiary     human

say-v1
  syn-struc
        1         root     say
                           cat      v
                  subj     root     $var1
                           cat      n
                  obj      root     $var2
                           cat      n
        2         root     say
                  cat      v
                  subj     root     $var1
                           cat      n
                  comp root         $var2
  sem-struc
        1 2       inform

                  agent    ^$var1
                  theme    ^$var2
                  instrument NL
```

Figure 3 example of a lexical entry

3.3 Mapping lexical syntactic-semantic information

The SDS building process is guided by the syntax-semantics interface manifested in the lexical syntactic and lexical semantic specification of lexical entries. So, one of the most key decisions in developing knowledgebase in OS is the specifications of ontological concept(s) for lexical entries.

Several mapping methods required according to different lexical cases:-

- direct mapping: when the semantics of the sense is fully described by a concept.
- modified: when no concept exactly matches the semantics of a sense, then take the closest in meaning and then modify some of its properties to construct a complex knowledge structure that quite accurately reflects the meaning of this sense.

Modified mappings are a powerful method for avoiding the proliferation of concepts in the ontology, the drawback being not only increased processing load, but also considerable acquisition work, since we have not yet found a way to automate it, unlike direct mapping, which can partially be automated.

3.4 Modality and Aspect:

Lexical units in the lexicon are assigned modality and aspect. Modality it is the attitude that the speaker holds towards the objects or situations which are in the propositional component in TMR.

In the present implementation of ontological semantics there are seven types of modality, used to express distinct speaker centered situational attitudes. For example “Epistemic” modality describes the speaker attitude toward the factivity of the proposition; its value is 1 if the speaker is assured of the reality of event, 0 in case of negation, and somewhere in between if the event is somewhat likely (the two instances of some corresponding to a scale measurement).

Aspect, which is mostly a feature of text meaning representation in ontological semantics, is currently implemented as a combination of one of the four PHASEes: begin, continue, end, and begin/continue/end .

4. Lexicon acquisition:

Acquisition is the lifeblood of ontological semantics. Through the acquisition process, trained acquirers describe the backbone of the knowledge format and template to this natural language processing approach.

But, the acquisition process can be difficult, redundant, and time-consuming, leading to a variety of errors and an ultimate slow-down of an already.

The acquisition process is not such a direct operation even for the master acquirer. Before proceeding in it several things should be considered

- the clearly stating of the domain and its related sub domains
- the availability of suitable source of information(corpus) that cover the whole aspect of a specific domain (Malaia in [3] implemented a source-topic variability matrix to cover a domain)
- how to specify the boundary between an ontological concept and a lexical item

For the acquisition process in OS three different layers of acquirers required. First layer is the master acquirer that formulates the initial knowledge units' templates, the second layer is the acquisition managers (highly trained acquirers) that manage the difference in features and classification of the templates and highly controls layer three. Layer three is the acquirers that usually perform the massive acquisition process using the ready made templates

And like all Knowledge-based applications which involve NLP, OS carried the stigma of being too expensive to develop, difficult to reuse as well as incapable of processing a broad range of inputs and this high price of development was due to the necessity to acquire all knowledge manually, using expensive expert-trained human acquirers.

The steps of lexical acquisition may be presented as follows:

1. polysemy reduction: decide how many senses for every word must be included into a lexicon entry: read the definitions of every word sense in a dictionary and try to merge as many senses as possible, so that a minimum number of senses remains;
2. syntactic description: describe the syntax of every sense of the word;
3. ontological matching: describe the semantics of every word sense by mapping it into an ontological concept, a property, a parameter value or any combination thereof;
4. adjusting lexical constraints: constrain the properties of the concept property or parameter, if necessary;
5. linking: link syntactic and semantic properties of a word sense.

One of the first and most important tools for acquirers is a clearly stated set of terms and accompanying definitions relevant to acquisition, centrally, the specification of the formats and of the semantics of the knowledge sources. Another, but extremely important tool for linguists hoping to successfully acquire lexical items in any particular domain, is a dictionary specific to the domain area. For example, when acquiring in the medical domain, researchers should use a medical dictionary, in the legal domain, a law dictionary, etc. Dictionaries are not only useful in providing definitions for humans in the ontology, housed on a centralized acquisition tool such as Purdue University's KBAE, but also in polysemy reduction, one of the major areas of focus for master acquirers.

Once it is there the development of a toolkit for acquisition can start. The toolkit includes acquisition interfaces, statistical corpus processing tools, a set of text corpora, a set of machine-readable dictionaries (MRDs), a suite of pedagogical tools (knowledge source descriptions, an acquisition tutorial, a help facility) and a database management system to maintain the data acquired (see Figure 4).

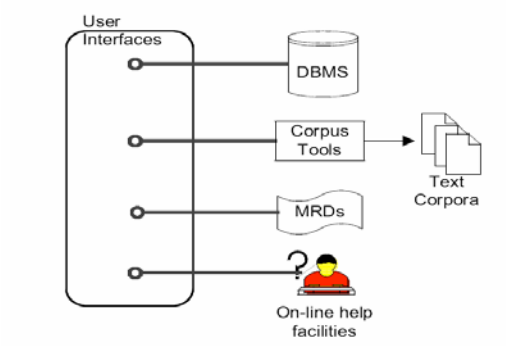

Figure 4. An ontological semantics acquisition toolkit

The various methodologies developed for acquisition of the static knowledge sources (the ontology, monolingual lexicons, onomasticons, and fact database) of ontological semantics necessarily involve, at this stage of development, considerable human participation, although the aim is to fully automate all processes involved in the approach, both in terms of acquisition and runtime procedures[1]. But this is till now unfortunately unavailable, although OS authors claim that OS in its current state utilizes every possible automation that could be done with in the current environment of acquisition, but all this done under the control of human acquirer.

4.1 Automatic lexicon acquisition

Since the lexicon is the main concern here, it will be worth it if we investigate the possible methods of automation that can be used for its acquisition.

The principle of complete coverage, to which ontological semantics is committed means that every sense of every lexical item should receive a lexical entry . “Every” in this context means every word or phrase sense in a corpus on which an application is based.

When acquiring a lexical entry, the most difficult part of the work is determining what concept(s) to use as the basis for the specification of the meaning of a lexical unit; the moment such a decision is made, the nature of the work becomes essentially determining which of the attributes values of a lexica entry to modify to fit the meaning of the lexeme.

The acquisition of lexical entries suitable for a specific domain can be done semi automatically through the use of set of techniques like rapid propagation and lexical rules.

- Rapid Propagation: In this procedure the master acquirer will build a sample entry for each specific class of lexemes. In this case this ready template can be used to facilitate the acquisition process of the lexical entries so that all what the acquirers have to do is copying this template to acquire the set of words in a specific class plus some minor modifications according to different words features. Some of the classes is of a small no of entries but in spite of that, the obvious benefit of using a ready made template for speedy and uniform acquisition of items in a class. And some such classes are quite large. One example of a large lexical class (over 250 members) whose acquisition can be rapidly propagated is that of the English adjectives of size.
- Lexical Rules: It finds economies in automatic propagation of lexicon entries on the basis of systematic relationships between classes of lexical entries, e.g., between verbs, such as abhor and corresponding deverbal adjective *abhorrent.* LRs consist of a left-hand side (LHS) which constrains the lexical entries to which the rule can apply and a right-hand side (RHS) which stipulates how the new lexical entry will differ from the original. Lexical entries which are produced by a LR are themselves eligible to match the LHS of an LR. Both sides of the LR can reference any zone of the

lexical entry; typically the RHS modifies the local syntactic information and the lexical semantic specification (or at least the syntax-semantic interface).

Inheritance is one type of cross-indexing used, for example, to indicate that a particular lexeme is of syntactic class, thus avoiding the need for a syntactic specification or syntactic features to be specified locally in the corresponding entry: the information will be inherited from the specification in the definition of the class. The same way can be used to inherent the semantic features of a set of lexemes in the same class.

But the limitations from the semantic side are much more than the syntactic side. Words with same classes may have different meaning attributes and hence different semantic features.

But from a theoretical point of view if the words can be classified according to their syntactic features, then there should be some way in such that words can be clustered according to it semantic features. But this will be on the account of the precise specification of semantic features for lexemes, which as against the attitude of the OS implementation.

> "ontological semantics would tend to produce complicated entries in the lexicon rather than in the ontology, and to this effect it provides lexicon acquisition with more expressive means and looser metasyntactic restrictions than the ontology."[5]

In OS theory entire stories may exist in the lexicon entries for the sake of clarity in the building process of TMR.

The assigning process of the concepts to a lexeme (which is the difficult step in lexeme acquisition) is done manually. And it is very crucial matter since it guides the SDS creation. As mentioned before, this can be done through direct and modified mapping. Depending on the semantic features of a lexemes mentioned in a dictionary by an inventory of case roles assigned to the syntactic structures of the lexemes, beside a specific domain corpus with an ontology, the semantic image of the lexeme can be reflected and matched against the ontology concepts to suite a no of concepts that are most appropriate to the lexeme in hand. Then depending on matching of syntactic and semantic image a filtration process can be made. This will not assure a successful indication of the right concept exactly but will limit the no of attested concepts to find the suitable concept(s). But again the formalism here is missing and the demand of a rule for representation and for assigning the case roles for specific entries still done manually. Besides that it is a costly way in terms of semantic representation of the lexical units in the dictionary. But it is shows that the automatic extraction of suitable concepts for lexical entries is visible when the lexical entries are of less complications and more general in their formalism (at least for the same classes).

This assumption indicates that more careful design and implementation of lexical entries are needs to be considered. The using of inheritance for the semantic features may provide sufficient amount of information with low cost. The manual intervention can not be avoided in the current state.

5. Expanding OS:

Two important issues is a matter of debate for the environment of OS. First the cost of the knowledge acquisition process of the resources for new languages. Second the portability of OS enterprise to new applications.

For the first issue Raskin et. al. in[9] claims that "The cost of the ontological semantic enterprise ran into the

millions at the peak of its development but, once expended, it does not have to be considered again", this is true for the ontology as it is a language independent source that represent the world knowledge around a specific domain. But it is not true for the lexicon which is language dependent source of information where specific language words and their properties are recorded. The acquisition process of new natural language lexicon is costly in terms of the availability of the experienced linguistic people that able to build the suitable formulation of the adequate predefined templates that lead the acquisition process. Also it costs in terms of rebuilding all these templates to match the special requirements of the new NL.

For the second issue, the portability of OS enterprise to new applications, it can be looked at from two different prospective.

1. For already implemented projects: in which the processes are already defined and proceeds in a specific ways. Such systems require redefinition of tasks and processes according to the format of TMR produced by OS system, which seems to be in adequate procedure to be implemented (only in certain cases).
2. For new projects: The format of the TMR which is the main result of OS system needs to be clearly stated and identified in a unique format and in a fixed way so that the processes in these systems can be build upon it.

Another important case is to be considered in the second issue is the specific demands for some applications. For example in machine translation, what is required from the translation in most cases is the input text as its exist in the source language.

6. Conclusions:

As a knowledge based system OS requires a vast amount of information regarding the world around a specific domain of application. The acquisition in its current state of these knowledge sources is done mainly manually, which limits the advantage of such leading comprehensive theory.

The detailed description of knowledge sources that are required for the processing in OS is the source of the problem. So a more formalism is needed to represent the knowledge to be make it possible to automate parts of the acquisition process more than what is currently exist. Lexicons in OS need to be more considered or more formal representations to be eligible for more automation processes.

OS is evolving continuously as new systems implementing it in response to needs for enhanced coverage and utility. Historically, a number of research projects have contributed to bring OS into its current state.

References :

[1] A. M. Ortiz, V. Raskin, and S. Nirenburg, "New developments in Ontological Semantics".

[2] B. A. Onyshkerych, "An ontological semantic framework for text analysis", Ph.D. thesis ,Pittsburgh , CMU-LT1-97-148, Carneyie Mellon University, pa 15213, May 1997.

[3] E. Malaia, "Digital identity management in ontological semantics: methodology and practice of domain", Ph.D. thesis, CERIAS Tech Report, Purdue University, USA, 2005.

[4] J. Spartz, E. Malaia, and C. Falk, "Methodology and tools for ontological semantic acguisition, CERIAS, Purdue Unv., USA, 2005.

[5] S. Nirenburg, and V. Raskin, "Ontological semantics", MIT Press, Cambirdage, MA, 2004.

[6] S. Nirenburg, S. Beale, and M. Mcshane, "Evaluating the performance of the Ontosem semantic analyzer", Proceedings of the ACL Workshop on Text Meaning Representation, 2004.

[7] S. Nirenburg, M. Mcshare, and S. Beale, "operative strategies in ontological semantics", Proceedings of HLT_NAACL-03 workshop on text meaning, Edmonton, Alberta, Canada, June,2003.

[8] M. Mcshare, S. Beale, and S. Nirenburg, "some meaning procedures of Ontological Semantics", Proceedings of LREC, 2004.

[9] V. Raskin, K. Triezenberg, E. Malaia, and O. Kranchina, "Ontological Semantic support for a specific domain "CERIAS, Purdue University, USA, 2005.