Lexical Knowledge Engineering: *MikroKosmos* Revisited

J. E. LONERGAN
[jjulialon@hotmail.com]

Department of Curriculum and Instruction & Department of Computer Science
New Mexico State University, Las Cruces, New Mexico, USA 88001

Abstract

This paper will describe the *MikroKosmos* methodology for the knowledge engineering of a computational lexicon used for text analysis. To do so, the paper outlines the general requirements for a knowledge base to be used for NLP, followed by specific requirements for building the lexical knowledge source. To highlight the issue of efficiency and reusability, the paper will contrast knowledge engineering for text analysis using the syntactic realizations of predicate logic against knowledge engineering for text analysis using lexical constraint satisfaction in combination with a world-model, or ontology. The paper suggests that the most efficient approach to the representation of natural language for analysis is the ontological and lexical semantic modeling (Onyshkevych & Nirenburg, 1994). Lexical semantics makes constraints accessible centrally from the concepts rather than through assertions of logic. References are made in this paper to the structuring of the English generation lexicon for the *MikroKosmos* KBMT project, a knowledge-based machine translation system based on deep-meaning analysis and text generation, developed at New Mexico State University, Computing Research Laboratory.[1]

Keywords

Knowledge engineering, semantic analyzer, *Mikrokosmos* Project.

1 Determining Meaning from Text

The ultimate goal of knowledge engineering for text analysis is to create the linguistic and computational structures intended for usage in extracting and representing meaning from a given text. One habitually assumed approach to processing text in NLP is the syntactic dependency or predicate logic approach. The notion of predicate lies on the binary semantic opposition between "theme/rheme" or "subject/ predicate" relationships [6]. From the syntactic perspective, a predicate can be said to be a named n-ary relation between arguments, which concerns itself more with capturing the semantics of events (verbs) or states rather than with the semantics of objects (nouns) [6]. The notion of syntactic predicate is equivalent to representing verbs and relational nouns, prepositions, and adjectives viewed together in a syntactic dependency frame.

The knowledge source built for predicate logics' syntactic dependency formalisms simply adorns the leaves of a phrase structure tree (taxonomy) with syntactic objects, where predicates are not in any way connected. Relationships between concepts are presented in the form of logic axioms (functions) that only cover directly specified relations [3]. The search space is the entire set of binary predicates, not just those that link the concept to another concept. This method of text analysis answers queries that are truth-conditional, all links in the syntactic pattern need to be known beforehand, and the constraints must be met literally. There are only constraints on arguments of predicates. This causes constraints to be either "too tight" or "too loose" for NLP purposes. Increasing the size of the knowledge base creates an imbalance with the existing control technique and it becomes unable to handle the explosion of information [3].

Predication structures represent lexical knowledge by providing a mapping from a word to a concept, not to the "lexical semantics" of the word. It is difficult to determine the meaning of a word by just the analyzing relations among lexical items at the surface level. A determination of semantic affinity requires a measure of conceptual relatedness. One salient requirement of a semantic network is the ability to search and find all the links from one concept to other concepts. Disambiguation is determined by the actual distance between word senses. Symbolic, object-oriented representations, engineered in a network of frames, arc labels

[1] http:://crl.nsmu.edu/Research/Projects/mikro for further information

(relations between concepts), and nodes (concepts) in an ontology, are designed for finding such relationships. The interplay of constraints is a major factor in determining the best overall semantic analysis. The knowledge source designed for determining meaning from syntactic dependencies does not rely on this type of interwoven semantic network, and therefore there is no method for comparing different choices and selecting the best combination of senses for a sentence. Constraints often do not have "yes" or "no" answers. Constraints are only tendencies, and metonymy and figurative language often override these tendencies [1].

Without the conceptual relatedness in a well-developed ontology, there is not enough knowledge available for disambiguation. Words in the same syntactic category can be expressed at very similar concepts, yet they can not be used interchangeably without making the sentence ungrammatical. With the simple provision of a rough mapping from word to concept with denotation predicate, there no clear separation between knowledge of English words in the lexicon and knowledge of concepts. It is very difficult and expensive to build knowledge without the necessary constraints. The knowledge base representation must be *object-oriented*. There must be a distinction between natural language (verbs and nouns) and the language of representation (events and objects). Meanings in the source text, not just its syntax, must be mapped into a set of interlingual symbols (ontology).

The semantic lexicon stands in contrast to the knowledge structure based on capturing meaning from syntactic dependencies and interactions at the word level, where verbs, nouns, prepositions, adjectives are together related with the syntactic dependencies. The semantic lexicon is constructed on the assumption that sentences can be broken into independent words of self-contained meaning [5, 6]. Knowledge units (lexemes) are lexicalized as *events* and *objects*, not as syntactically related concepts [6]. Concepts are not equivalent to the syntactic predicate because they have no surface realization at all [6]. A semantic lexicon must be acquired in close connection with the specifications of the world model in a semantic network (an ontology), the frames for representing text meaning (language-independent schemata), and the computational semantic search engine [5]. The lexicon and the conceptual inheritance structure both must undergo algorithmic operations. The semantic lexicon contains generative devices (argument, event type, speaker attitude, aspectual and modal information, and semantic relations) that define the semantic relatedness between syntactically distinct expressions and guide constraint-based search through a complex tree [1, 2].

2 The Semantic Network

The ontology is the knowledge resource that contains concepts and well-defined attributes and relationships with other concepts [4]. The ontology defines for each semantic concept the set of arcs that are allowed/expected, as well as the appropriate filler concepts. Concepts are primitive symbols of a world model which includes *objects*, *events* and *properties* organized in a complex hierarchy of language-independent concepts. The concepts do not refer to the actual real word entity, but to the generic notion of entity. The concepts are constructed following super ordinates, or hyponymy relations (IS-A links). And each hyponym inherits all the features of the more generic concept and then adds at least one feature that distinguishes it from its super-ordinate and from any other hyponym of that super-ordinate [11]. In addition, its organization into a taxonomy via IS-A links, the ontology should contain numerous other links between concepts. A general requirement for a knowledge base is a large amount of broad-coverage world knowledge: taxonomic knowledge, attributes, relations between concepts and constraints on them, instances, and lexical knowledge [3, 4]. A link between two concepts is a *property*. A *property* is a *relation* which has a pointer to the destination of the links (DOMAIN). Each concept node has a value, or domain. Unary constraints restrict the value of domain without reference to any other variable; and binary constraints restrict the values of a variable by comparing it to another variable [1]. This ontology creates a lexical inheritance system of concepts where there are no uninterrupted (unconnected) symbols. In computational semantics, unary constraints correspond to selecting the appropriate word-senses from the lexicon based on the word used and the surrounding syntax [1].

Ontological concepts can be instantiated, or a specific instance of a concept is produced to signify a particular mention of this lexeme in a text during processing. An ontology that will actually be used in application will include such *properties* as SUBJECT, and SUBJECT-OF which IS-A EVENT-OBJECT-RELATION with [domain EVENT OBJECT range EVENT OBJECT], as well as semantic dependency relations that have been traditionally referred to as *case roles*, such as AGENT and AGENT-OF, which IS-A INVERSE-CASE-ROLE-RELATION with [domain HUMAN range EVENT], or THEME which IS-A CASE-ROLE-RELATION with [domain OBJECT EVENT range OBJECT EVENT] [5]. There is always an inverse relation. Ontological concepts are represented as frames, and properties are slots in the frame. Concepts are represented as nodes, and properties are links between nodes. Each of the variables have binary constraints. Each variable can take on a node-consistency to pick the correct lexical entry based on surrounding syntax, and arch-consistency, where a value in each domain must satisfy the binary constraint on that arc [1]. The constraints are contained in the lexicon, and are applied before beginning the search algorithm. Thus, the knowledge-base acts as a hierarchy that can be searched upward or downward by powerful search mechanisms.

The knowledge-base model solves for acquire subjectivity because it contains information that can be used to directly model the meaning of external words. The knowledge is represented with English-labels or values. The labels are language-independent. The meaning is derived from the location of the lexical item in the taxonomy, and from the different relations that occur between it and other concepts such as: IS-A links/hierarchical relations, linguistic case roles/constraint satisfactions, attributes and property relations [2].

3 Knowledge Engineering in the Semantic Lexicon

The lexicon is a dictionary that contains zones of information specifying declarative knowledge about the world [9]. Its entries are based on fine-grain semantic distinctions. The information is stored in zones. There are three important zones in the *MikroKosmos* English-generation lexicon: the SYNtactic zone which specifies sub categorization patterns, the SEMantic zone which links the lexicon to the ontology, and the SYN-SEM zone, which links the sub categorization patterns with the semantic arguments. The lexicalized syntactic information it contains is used in syntactic parsing and in establishing syntax-semantic structure mappings [2, 5, 7, 10].

The SEMstruc zone is the concept specification, which is based on semantic meaning. It is also an instruction to the semantic analyzer to add an instance of the ontological concept in question to the text meaning representation [1, 5]. The SEM zone is the zone that connects the lexicon with the ontology becoming the locus of the atomic links between lexical units in text and the language-neutral meaning representation, or text meaning representation [5]. The SEM zone connects the lexicon with a preexisting, well-structured, and information rich knowledge base, or ontology. The lexicon entries are mapped into this knowledge base[9].

The syntactic information is stored in the lexicon at SYNstruc zone[2, 5, 7, 10]. This zone is an under-specified piece of a parse tree. This syntactic-structure zone is a fs-pattern that specifies the arguments for ROOTS, or the base form of lexemes. The building of the syntax of a lexeme entry is done by specifying all information related to the syntactic behavior of the word. This allows for encoding of the argument structure. The information specified for the SYN zone guides the search algorithms and creates ontological concept instantiations.

Input: *The judge repealed the laws on black segregation.*

LEXICON [constraint satisfaction through concept instantiations]

```
Judge-N1
       Sem: JUDGE
       Syn: LG-np

Repeal-V1
       Sem: REPEAL-LAW
       Syn: LG-np[AGENT]-v-
np[THEME]-pp_adj(opt)[on]-
ARG3[SUBJECT]

Law-N2
       Sem: LAW
       Syn: LG-np-(opt)oblique[on
       about]SUBJECT

Black-Adj1
       Sem: HUMAN is-a ANIMATE
              HAS-ORIGIN
                     DOMAIN: HUMAN
                     RANGE: AFRICA
       Syn: LG-adj-att-pred
(attributive and predicative)

Segregation-N3
       Sem: SEGREGATE
```

```
        Syn: LG-
np_oblique1(opt)[of][THEME]-
oblique2(opt)[by][AGENT]
```

The SYN-SEM zone links the elements between the SYN-zone and the SEM-zone. Mappings between the syntactic complements and the semantic roles are defined in the example below.

```
(2)
repeal-V1,
        SYN: root [0]
                subj [1]: [cat: NP]
                obj [2]: [cat:  NP]
        SEM: REPEAL-LAW
        agent: [11]human
        theme: [21]legal-object

        SEM: SUBJECT
        domain: [12]legal-object
        range: [14] segregate

SYN-SEM link: root
subj [1]: cat: NP; sem:[11] human
obj [2]: cat: NP;  sem: [21] legal-
object
arg3 [3]: cat: NP; sem: [14]
segregate
```

Instantiated concepts form circles of interdependence. This is accomplished with the Hunter-Gatherer Algorithm [1], where branch-and-bound circuits work at a "higher level" of interaction than predicate calculus, higher than interactions at the word level, where verbs, nouns, prepositions, adjectives are together related with the syntactic dependency [6]. Here instead the process produces a text meaning representation, saturated with information "in the blanks". Language independent schematas in the *Mikrokosmos* project are called TMRs. It is the ontology that supplies the world knowledge. The semantic and syntactic patterns associated with each unit in the lexicon act as *object-oriented* constraints on the ontology to produce a meaning representation of a given input. The lexicon is intimately connected with an ontological model of the world and with the text meaning representation language.

4 Language Independent Schemata

Interlingual language schemata are frame-based and result from the syntactic parse of the input. The analyzer collects all possible lexicon entries and examines the zones of the lexicon that deal with syntax-semantics mappings in order to construct a list of constraints that must be satisfied for that word sense. The best combination is selected by determining how well a given sense of a word combines with sense of other words in the sentence to form a coherent meaning for the entire text [1, 3]. This knowledge for this analysis is stored in the lexicon. The lexical zones are the building blocks for the language independent schemata, which are thus a dynamic construct [4]. The instantiated schemata are combined in well-defined ways to analyze input text and to produce and structure lexemes for the generator lexicon [9].

The Judge repealed the law on black segregation.

The semantic analyzer [1, 2] focuses on selection constraints in the lexicon and the ontology for each pair of syntactically dependent words, the desired meaning can be interpreted and represented interlingually in a TMR as follows:

```
Repeal-1
        Agent: Judge-1
        Theme: Legal-Object-1

Judge-1
        Agent-of: Repeal-1

Legal-Object-1
        Theme-of: Repeal-1
        Subject: Segregate-1

Segregate-1
        Subject-of: Legal-Object-1
        Theme: *Dummy[human]-2

Origin-1
        Domain: *Dummy[human]-2
        Range: Africa-1

*Dummy[human]-2
        Has-Origin: Africa-1
        Theme-of: Segregate-1
```

The chosen word senses are assembled into frames using lexical semantic representations from the lexicon, as shown in example above. The Dummy role is a separate sub-plan that must be made for the *theme* relation. For each semantic concept and relations that is included in the lexicon entry, a dummy sub-plan is created [1]. The additional sub-plan is needed to relate the unspecified *theme* of the segregate-event to attribute describing its type. The sub-plan corresponds to the missing relation. The dummy receives a constraint that may be used only if needed. The main benefit this gives is that a stable set of "variables" can be created.

The ontology is neutral and the lexicon is language specific. The latter is mapped into the former. And because the structure of the frames are under specified with respect to other languages, each source language can map to the ontology. Language independent schemas are grounded in the ontology, the ontology provides the search space for the powerful search mechanisms directed by rules in the lexicon, and a target language English generator produces off of the instantiations, a target language. The ontology is the common ground between the analyzer lexicons and the generator lexicon, thus a translator between different languages.

5 Conclusions

This paper deals with the issue of how to efficiently capture the meaning of lexical elements from a given text. The paper described how to best represent the meaning, given ontological and lexical modeling in the *MikroKosmos* project. For *Mikrokosmos*, text generation and semantic analysis are equivalent. The "variables" are words, and the "plans" are word senses. The analyzer tries to plan the combination of word senses that best describes the semantics of the input. Predicate functions cannot overcome the lack of structure in representation by providing a mapping from an external word to a concept, not to the "lexical semantics" of the word. With predicate logic there exists the inability of the nodes to communicate with each other. Real-world problems have complex semantics, and constraints are not "yes" or "no." Also, the distinction between natural language (verbs and nouns) and the language of representation (events and objects) is still not widely appreciated by knowledge base engineers. Reusability can be attained when there is a separation of knowledge into separate lexical, ontology, and fact databases, build without the syntactic predicate logic that is the habitually assumed as a search method. Rules that keep the conceptual knowledge and linguistic knowledge together make it difficult to extend the use of the knowledge sources for processing languages other than English. And as, psycholinguistics posits, human sentence processing has a strong lexical basis with realistic grammars that are systems of flexible constraints, not shallow structural manipulations.

References

1. S. Beale, Hunter-Gatherer: Applying Constraint Satisfaction, Branch-and-Bound and Solution Synthesis to Computational Semantics, PhD Dissertation, School of Computer Science, Carnegie Mellon University, CRL Technical Report, MCCS-96-289, NMSU, 1996.
2. S. Beale, S. Nirenburg, K. Mahesh, Semantic Analysis in the Mikrokosmos Machine Translation Project. In Proceedings Symposium on NLP, Kaset Sart University, Bangkok, Thailand, 1995.
3. K. Mahesh., S. Nirenburg, J. Cowie, D. Farwell, An Assessment of Cyc for Natural Language Processing, CRL, Technical Report MCCS-96-302, New Mexico State University, 1996.
4. K. Mahesh, Ontology Development for Machine Translation: Ideology and Methodology, CRL, Technical Report MCCS-96-292, New Mexico State University, 1996.
5. B. Onyshkevych and S. Nirenburg, A Lexicon for Knowledge-Based MT, PhD dissertation, U.S. Department of Defense and Carnegie Mellon University, 1994.
6. E. Viegas, K. Mahesh, S. Nirenburg and S. Beale, *Semantics in Action*. The Netherlands: Kluwer Academy Press, 1999
7. E. Viegas and V. Raskin, Computational Semantic Lexicon Acquisition: Methodology and Guidelines, CRL, Technical Report MCCS-98-315, New Mexico State University, 1998.
8. E. Viegas, and P. Saint-Dizier (Eds.) *Computational Lexical Semantics*. Cambridge Univ. Press, 1995.
9. E. Viegas, A. Ruelas, J. Lonergan, J. Longwell, S. Beale, and S. Nirenburg. Developing a Large-Scale Semantic LKB to Suit an Intelligent Planner, In proceedings of, 7th European Workshop on Natural Language Processing, Toulouse, France, 1999.
10. R. Zajac, and E. Viegas, The Generic Structure of a lexical knowledge base entry. Computing Research Laboratory, Technical Report, New Mexico State University, 1998.
11. G. Miller, Nouns in WordNet: A Lexical Inheritance System. In International Journal of Lexicography 3.4, 1990.