# Natural Langauge Processing and Knowledge Representation

Lucja M. Iwanska and Stuart C. Shapiro

January 19, 2011

## 1 Natural Language is a Powerful Knowledge Representation System: The UNO Model

Conjecture: Natural language as a knowledge representation system. This model is fully implemented and can handle a corpora of thousands of documents.

**Expressive and computationally tractable** NL has a rich structure of complex boolean expressions (negation, conjuction, disjuction, adjectival/adverbial modification). This structure lends itself to easily-computable inferences. Close relationship between syntax and structure.

**General purpose** Uniformity of representation and reasoning.

**Logical contradiction and logical redundancy** Serve to identify knowledge gaps and convey nonliteral meanings.

**Context dependency** Reflects process of knowledge acquisition.

**Facilitates machine learning** Provides an expressive mechanism for formalization and easily parsable constructions to convey taxonomic language.

**Mixes object- and meta-level descriptions** Same representational and inferential mechanism to draw inferences about the environment (object) and to reason about its own knowledge (meta). Drawback: paradoxes.

### 1.1 Conjecture Validation

Balancing theoretical research with practical implementation involving shortcuts and hacks. Can anything be proven about natural language?

### 1.1.1 Representational and Inferential Strengths and Weaknesses

NL can be viewed as a formal representational language, more than just an interface.

The representation of language does not have to mean a non-natural data scheme. By representing language as language, we can build a system the handles the meaning of langauge.

Knowledge representation schemes not developed explicitly for the purposes of understanding natural language tend to be very removed from the language. This leads to weak and *wrong* representations. Formal systems tend to be built for artificial data rather than real-world corpora. However, being different from NL does not have to mean worse, as this distance can reveal weaknesses in the representational abilities of NL itself, constituting a representational alternative. That said, only narror, specialized representation systems present this possibility. Formal systems do not currently approach a general-purpose capability.

### 1.1.2 UNO Model of Natural Language

UNO offers a solid computational and mathematical framework intact with linguistic theories. The model constantly updates its knowledge base *and* automates inferencing by the same semantically clean computational mechanism of performing boolean operations on the representation of natural langauge input and the representation within the knowledge base (existing knowledge). UNO closesly mimics the structure and capabilities of natural language, which allows UNO to build a knowledge base from an existing corpora of text.

NLP as Knowledge Acquisition. View natural language as speaker/hearer-based, rather than solely speaker (taking context of existing knowledge of the hearer into account). Could be crucial in biulding the knowledge graph.

"Pure" Meaning in Natural Language: A sentence that any native speaker could understand without any external knowledge. Sources of universally shared knowledge are the semantics, mathematics and computability of: 1) generalized quantifiers, 2) adjectival/adverbial modification, 3) boolean modification, 4) underspecificity of lexically simple scalars.

### 1.1.3 Research Motivation behing this Model

Theory: Show the computational aspects of natural language. Engineering: Demonstrate that this solution scales up and works on large, real-life corpora.

Formal Theories and Implementation: Provide a complete framework for computing literal meanigns of natural langauge and account for nonliteral meaning.

- A general semantic model of negation in natural language. Involves representation and inference for sentences and tiny texts involving explicit negation at different

syntactic levels and accounting for the complexity between negation, conjuction, disjunction, quantifiers, adjectival/adverbial modification and scalar expressions.

- Temporal logic: reasoning about relative and absolute time in natural language. First done on newspaper articles.
- Semantics, pragmatics and context of intensional negative adjectives. Infer meaning of sentences modified by adjectives like "alleged" and "toy."

Engineering solutions: "Weak" methods. Problems:

- Automated processing of narratives written by grade students. Involves preferential sentence-level parsing and extracting prepositional phrases.
- Discourse processing. Computing structures with nonlinearly distributed knowledge.
- Extraction of names, numbers, locations, dates, etc. from a large volume of newspaper articles.

#### 1.1.4 Technical Aspects of UNO

- Sentences asserting properties ("John is neither a good nor hard-working nurse") are represented by the following equation:
  $type == \{< P_1, TP_1 >, < P_2, TP_2 >, \ldots, < P_n, TP_n >\}$ where *type* is the UNO representation of a noun phrase or name of concept, $P$ is a property value, and $TP$ is a set $< t, p >$ such that property $P$ holds at temporal interval $t$ with the probability $p$. Both $t$ and $p$ are UNO representations of natural language expressions that describe temporal and probabilistic information.
- UNO uses its knowledge base bi-directionally for both answering questions about the properties of a particular entity and matching given properties against the properties of a known entity or concept

(When building a knowledge graph, how do I avoid redundancy? Make sure the retrieval system is robust/flexible enough to prevent redundant vertices, leading to lower-than-expected context scores)

Underspecified terms: building blocks of UNO representations. NL is very underspecified—reliance on context is crucial. In parallel, UNO models are also underspecified, allowing context to be taken into account from the knowledge base. If a sentence lacks temporal or probablistic information, the model will equivalently lack such information.

If a property holds for a single temporal interval or a set of intervals which can be described via a single temporal expression, the set notation can be flattened to $< P, t, p >$.

Building blocks of UNO: sets $[a_1, a_2, \ldots, a_n]$ whose elements $a_i$ are terms. These terms are record- and graph-like structures consisting of two elements: 1) a *head* (type) and 2) a

*body*, a list of attribute-value pairs $attribute => value$ where attributes are symbols and values are sets of $|n > 0|$. Example:

$[woman(health \rightarrow sick, happy \rightarrow (\not happy)(degree \rightarrow very))]$