

# Introduction to Knowledge Resources

Donna K. Byron

## 1 Overview

This course is in line with the general problem explored in AI: making sense out of observables through the clever application of background knowledge. Sometimes that background knowledge can be represented as collections of probabilities learned from exposure to data, or weights in a network, etc. In this case, the support for reasoning comes from the background knowledge that is cast as a collection of logical statements in some form of predicate calculus. Our focus in this seminar is to find ways to automatically populate this knowledge base of logical statements.

## 2 Predicate Calc basics

Knowledge representation is the formal language we will use to build a model of the world of interest for our problem. We must define the inventory of symbols that will be used, and represent the facts about the world using these symbols in a way that can be subjected to reliable reasoning techniques.

- The set of objects  $D$  about which knowledge is being expressed is called the *Universe of discourse*
- individuals are represented by constants, i.e. DonnaByron or ColumbusOhio or Horse746638
- Formally, a *conceptualization* is a triple with the following elements: 1) the universe of discourse, 2) the set of functional mappings, 3) the set of relations between objects
- predicates are names for relations or functions. They are composed of the objects in  $D$  that have that relation. Think of them as concepts or word senses (look at “house” on wordnet)
  - different senses can be attached to the same lexical item, but different inferences should be made depending on the sense
  - relational predicates have an arity, and argument positions (see (KP02) for a good discussion of argument structure)
  - A predicate and its arguments form the basic kind of well-formed-statement in PC, i.e.  $\exists x \text{Dog}(x)$  or  $\text{Dog}(\text{FREDDY})$  and they can be used to construct more complex logical sentences  $\forall x \text{Dog}(x) \rightarrow \text{Mammal}(x)$
- organizing the type hierarchy into an ontology means that the objects can inherit properties. i.e. adding  $\forall x \text{Mammal}(x) \rightarrow \text{hasSkin}(x)$
- We would like the collection of sentences contained in a KB to be a true representation of the world, and to be as complete as is practical
- There are many different true representations of the world

A statement in natural language often contains many links to knowledge, and we would ideally like to be able to represent it all. In addition to the asserted content, other implications and presuppositions hold, which the speaker of the sentence is committed to believing, and which we might like to compute. For example, an agent that believes “John’s sister stopped doing yoga” is in an inconsistent state if it also believes that people do not do yoga, or that John doesn’t have a sister, or that John’s sister never did yoga.

## 3 What is knowledge used for in language processing

1. Text says: “The northern trout lives in cold mountain streams.”
  - Information retrieval should return this document when servicing a query for documents about fish.
  - A question answering system should be able to use this sentence to answer the question: “can fish live in cold water?”
2. Text says: “The plate with my sandwich on it was cracked, but I didn’t notice until I ate it.”
  - we want to calculate that ‘ate’ prefers sandwich over plate, even if we’ve never encountered the phrase ‘ate sandwich’ before.

- the fact that sandwiches can be eaten is useful both for interpreting this particular sentence, and as a matter of counting in training examples for future processing
  - two ways to do that.
    - (a) Either pull text about sandwiches and text about plates and see which ones are more like the typical things found in the phrase 'ate x'
    - (b) augment the text with hyponyms (sandwich is a type of food, plate is a type of dinnerware) and infer which is edible
3. cluster by topic through word disambiguation: ontology gives you the inventory of word senses
  4. The text says "I was looking at Saturn with my Zeiss scope"
    - we want to parse it so that the 'with' phrase modifies looking, not Saturn

#### 4 Existing Knowledge Resources

What kind of resources have shown themselves to be useful?

1. semantic frames, argument positions, verb subcategorizations (this can be tracked at either the lexical item or the word sense level)
2. graphs or networks of word relations:
  - node typically contains a set (synset, cluster, category)
  - relation names:
    - meronym (parts or pieces of a collection)
    - hyponym/hypernym (pull up wordnet and show some examples)

The cast of characters:

- wordnet (show all relations for avoid)
- framenet (click on lexical units, k keep away = avoid)
- conceptnet, open mind (at MIT media lab)
- cyc (go to documentation for geopoliticalentity)
- treebank <http://www.cis.upenn.edu/treebank/switch-samp-bkt.html>
  - Brown corpus: mixed text including fiction and textbooks
  - Wall Street Journal: news stories including ads and short announcements
  - Switchboard:

## 5 Language Processes

- parsing: structural analysis
  - partial parsing: <http://nltk.sourceforge.net/lite/doc/en/chunk.html>
  - full parsing: WSJ or PTB trees: bracketing
- stemming: removing affixes or standardizing out morphology to get the word to a base form (i.e. running/runner/ran → run)
- Updating the KB with new logical sentences is sometimes called 'incorporating' the knowledge (as a step in language processing)
- Bag-of-words methods take a document and turn it into an unordered list or context vector of word (lemma) counts.
- N-gram methods count the words in sequence as they appear in the document.
- Pattern-detection methods consider small sections of text of either surface word patterns or regular expressions, sometimes with chunk parsing. NPs can be arbitrarily complex. The simplest patterns ignore structure in the hope that, given enough data, simple language processing will suffice.

NEW YORK (AP) – An unfinished tale by J.R.R. Tolkien has been edited by his son into a completed work and will be released next spring, the U.S. and British publishers announced Monday.

Christopher Tolkien has spent the past 30 years working on "The Children of Hurin," an epic tale his father began in 1918 and later abandoned. Excerpts of "The Children of Hurin," which includes the elves and dwarves of Tolkien's "The Lord of the Rings" and other works, have been published before.

"It has seemed to me for a long time that there was a good case for presenting my father's long version of the legend of the 'Children of Hurin' as an independent work, between its own covers," Christopher Tolkien said in a statement.

The new book will be published by Houghton Mifflin in the United States and HarperCollins in England.

## 6 Evaluation

- supervised: The technique starts with labeled inputs like the ones that the technique is attempting to construct. Its output is judged on whether it can reproduce those labels when they are withheld in a portion of the data.
- unsupervised: The technique does not start with labels on the data of interest, but might have some labeled data from another source. In other words, you don't have to create new labeled data to begin the learning process.
- semi-supervised: The technique does not produce output that is used as is. It is sent to humans to judge and filter.

Supervised processes often invoke an initial step of quality control over the labeling process of the gold-standard answer keys, which you will see reported in the paper as the Inter-Annotator reliability score, or Kappa. The Kappa measures the actual agreement obtained among multiple labelers, compared to how much they would be expected to agree through random chance (where random chance here is normalized against the proportion of label occurrences in the corpus).

The computation of *Kappa* is:

P(A) is percent actual agreement

P(E) is the percent expected agreement

$$P(E) = \sum_{n=1}^k \frac{Count_n^2}{Total\ items}$$

k is the number of possible values of the labeled item

$$\kappa = \frac{P(A) - P(E)}{1 - P(A)}$$

## 7 Glossary

**Subcategorization Frame** : The syntactic carrier phrases that are possible to express arguments and modifiers for a specific predicate. For example, the verb RUN can allow modification with a duration value, e.g. “RUN for two hours”, or a destination “RUN to the campus”. Both use the subcat pattern VB PP. RUN does not appear in the pattern VB SCOMP (e.g. RUN that I told you”).

**Bootstrap** : (verb) to promote or develop by initiative and effort with little or no assistance

**Open/Closed** word classes: Open classes are productive parts of speech in which new words can be invented or imported: nouns, verbs, adjectives, etc. Closed classes are the parts of speech in which you have to stick to an existing set, such as pronouns, conjunctions, determiners.

**Stop word** : a closed-class word appear in a text, such as 'the' or 'it', which does not assist shallow interpretation in determining the semantic content of the text.

**Head** : The primary structural constituent of a grammatical phrase, such as the Noun in a noun phrase or a Verb in a verb phrase. Other portions of the phrase are considered to be dependent on, or modifiers of, the head.

**Stem** : The bare form of a word reduced of any affixes or inflexional morphology. This removes spelling differences picked up by the word when it is used as a noun, verb, adjective, or in different tenses, etc. Ex: Runs, runner, ran.

**tf/idf** : Term frequency divided by inverse document frequency. This is a commonly used measure that finds the words that distinguish one context from another, as opposed to frequent words that appear in all contexts.

**Precision and Recall** : Two measures of classification correctness, stated as percents. Consider a set  $N$  of items of size  $i$  which are to be labeled based on an inventory of class labels.  $A_i$  is the correct label for element  $i$ . The resulting Precision is the sum over  $i$  of elements for which  $A_i = f(N_i)$  divided by  $i$ . Recall is the count of items where  $f(N_i) = k$  divided by the true size of class  $k$ . When all  $i$  elements of  $N$  are assigned a label (forced choice), Precision = Recall, but precision and recall over individual classes can vary.

## References

Paul Kingsbury and Martha Palmer. From treebank to propbank. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-02)*, Las Palmas, Canary Islands, Spain, May 28 – June 3 2002.