# Lexical Inference and the Problem of the Long Tail

**David D. McDonald**
BBN Technologies
10 Moulton Street, Cambridge, MA 02138
dmcdonald@bbn.com

**Abstract**

No one's semantic lexicon is ever complete. People cope through the adroit use of context. Machines should do the same. In this position paper I suggest that a large part of this context consists of deep knowledge about adjectives, adverbials, and other linguistic forms whose meaning cuts across subject domains. This is a new direction for research that complements the large effort on ontology induction and could accelerate work on reading for reasoning.

## The long tail of words

Even a highly literate reader will occasionally encounter a word that they don't know or have only encountered inside a frozen idiom. Anyone with children will appreciate their ability to adapt well-known words to new purposes.

This is Zipf's law in action. There is no end to the low frequency tail of words, word-senses, and turns of phrase that one will encounter for the first time. Yet people are able to deal with these when they see them. We work out the gist of what they must mean by establishing what would be semantically and pragmatically consistent with the context.

Language understanding systems will have to do the same, especially in this era where dealing with massive corpora with minimal filtering is becoming the norm. How we do this depends on our goal. Do we want a deep understanding that can be put to use in complex problem solving, or are we content with skimming for lists of facts.

### Domain-specific, concepts-first NLP development

I approach this from a paradigm where the machine is reading for content that will be used by a live, continuously running KR&R system ('knowledge representation and reasoning'). In this approach, a good descriptive logical form such as MRS [Copestake et al. 2005] is just a starting point, not an end in itself.

From a technical perspective as I see it, the analysis system (parser et al.), under the direction of inference rules grounded in a knowledge base, is directly feeding a database that respects the uniqueness principle [Maida & Shapiro 1982]. This means that every entity in the universe of discourse has just one referent in the KR. There is only one object denoting the current 'financial crisis', and it is directly incorporated into all of the propositions about it; syntactic constituents in the parses refer to individuals or concepts in the KR&R system, not substrings of the text.

Such an approach works particularly well if the vocabulary for the lexicalized grammar is projected from the conceptual model [McDonald 1983]. This means that the model should be developed first – presently an expertise heavy, time consuming task that is done domain by domain, sublanguage by sublanguage.

Because many of us have done it before, it is easy to imagine that we can work up such a rich conceptual model – one that is thorough enough to support reasoning and prediction – along with a an accompanying lexicon for a particular content domain such as sectarian bombings in Iraq as reported by standard news organizations or executive retirements and job changes as reported by the WSJ. Of course that is a far cry from an 'internet-class' corpus, but I do not believe it is worth our time to even consider attempting working at that scale if the goal is a deep analysis of content that a program can reason with.

Nevertheless, we see the effect of the long tail whenever we extend our set of deep models to a new domain. Many of the words our system encounters will already be known,[2] but words that are specific to the new domain will have to be modeled. The question is how we accelerate this modeling process.

I propose that we invest our lexical semantics modeling effort in the content words that are common to virtually all domains: high frequency classes of adjective and adverbs, along with, of course, the major classes of adverbials such as time and location. In the rest of this paper I will sketch how this can be done – how we can use words like *major* or *such as* to construct and refine a semantic model of the unknown words they modify.

## Inference from Context

We are all familiar with Lewis Carroll's poem the Jabberwocky, where words that almost none of us know have



[2] By 'known' I mean that the lemma is grounded in the KR&R system that is 'listening' to what is being read. It has a mapping to a concept inside the listener that has a pragmatic consequence in the KR&R listener's situations and scripts.

very clear grammatical roles. We deduce their roles by making inferences from the function words and morphology (in bold).

> `**Twas** [it was] brillig**, and the** slith**y** tove**s did** gyre **and** gimble **in the** wabe: **all** mims**y were the** borogrove**s**, **and the** mome rath**s** outgrabe.

After reading this, we can answer questions like where the toves were gyring; we know that these toves were slithy and the borogroves mimsy; and that the mome raths are/were in the state/activity of outgrabe/ing/ness.

## Cross-domain semantic context

We can do the same thing in the semantic domain. We use knowledge about modifiers and other forms to impose constraints on the categorization of their heads. The modifiers constituent a context that allows us to position an unknown word in a subsumption lattice. As the same word is seen in the context of other deeply understood modifiers, this categorization can be refined: the concept we create for the word moved further down in the lattice. If multiple words are observed in the same suite of contexts (appear at the same or nearby point in the lattice), we can conjecture they may be synonyms, which is a good point for injecting human expertise into the process. Axioms or other rules of inference associated with the organizing concepts in the lattice establish their meaning within the KR&R.

The basic idea is to treat each (class of) modifier(s) as a predicate as it would be in MRS or Hobbs Normal Form [Friedland et al.]. This predicate will define value restrictions on its arguments(s), and will come with a set of inference rules. If we read *three swans*, we know that whatever else swans may be, they are something that can be counted; from *majority option* we can infer that there is a 'minority opinion' even if we haven't read about it yet. The space of categorizations that these value restrictions are drawn from forms our subsumption lattice. Care in its design and the design of the inferences pays dividends in the leverage the context provides.

To get a sense of proportion, lets us look at this paragraph chosen at random[1] from the New York Times. Cross-domain modifiers are in *red italics*; domain-specific terms are underlined in blue; broad domain terms in bold **green**.

> *In* **Asia**, stocks *had already* closed *lower*. To *quell* **fears** *before the* open*ing* of **European** markets, the Fed and *other* central banks **announced** they would *make* **$180 billion** *available*, *in an effort to* get banks *to start* lend*ing* to *each other again*. The Fed *had agreed to* open its discount window to *make* loans *available to* money market funds *to prevent further* runs.

[1] This is the paragraph my eyes first lit on when I looked up at my browser after deciding I needed a new example. www.nytimes.com/2008/10/02/business/02crisis.html.

Discounting the function words and the broadly applicable classes of content words and phrases (countries, announcements by spokesmen), half of the terms are specific to the domain (*stocks*, *discount window*) and will only make sense to a KR&R program that has a model for them. The other half of them are terms whose meaning does not depend on the domain (*had already*, *quell*, *before the Xing*). We can use the inferences associated with these domain-independent terms to begin to pin down the meaning of the unknown, domain-specific terms.

This is a position paper, reflecting the fact that I have programmed only a few examples while awaiting funding. But let me give a superficial sense of what this cross-domain semantic context lets us conclude. From *had already Xed* we know that 'stocks' refers to an eventuality, either directly or metonymically. From the fact that 'stocks' *verbed lower*, we know that it is scalar. These are cases where category constraints associated with the terms let us ascribe properties to the concept we have formed to stand for the meaning of the unknown (to the program) head word *stocks*. These properties are the basis for its position in the lattice that forms the backbone of our ontology.

More examples. Because the 'fears' have to be *quelled* we know they already exist. The use of *other* lets us infer that 'the Fed' and 'central backs' are of the same kind. We can ascribe the ability to 'lend' to 'banks' because they may *start* to do it. From *again* we infer that they used to do it. From *each other* we know that 'banks' 'loan' to 'banks'. (If this sounds odd because you can't turn off your own knowledge of finance, try replacing the heads with head words from Jabberwocky.)

This enterprise is not without pitfalls. Candidates for context-defining terms must be thoroughly explored and vetted. Word sense differences are an issue (*bank*). But if we are careful about our sources (straight news, not quotes from politicians), we get qualitative categorial judgments about the words, something that linguists, particularly cognitive linguists, have done for decades [Talmy 2000]; we need to do the same computationally.

## References


Copestake, Ann, Dan Flickinger, Carl Pollard, Ivan Sag. (2005) "Minimal Recursion Semantics: An Introduction". Research on Language and Computation 3:281-332.

Friedland, Noah, David Israel, Peter Clark, Eduard Hovy, Jerry Hobbs, Rutu Mulkar, Bruce Porter, Ken Barker, Ralph Weischedel, Paul Martin. "Project Mobius: a study on the feasibility of Learning by Reading. www.noahfriedland.com/uploads/Moebius_Final_Report.pdf Downloaded 10/2/2008.

Maida, Anthony & Stuart Shapiro (1982), "Intensional Concepts in Propositional Semantic Networks", *Cognitive Science* **6**, pp. 291-330

McDonald, David. (1993) "The Interplay of Syntactic and Semantic Node Labels in Partial Parsing". Proc. Third International Workshop on Parsing Technologies, August 10-13, 1993 Tilburg, The Netherlands, pp. 171-186; revised version in Bunt and To-

mita (eds.), Recent Advances in Parsing Technology, Kluwer Academic Publishers, pgs. 295-323.
Talmy, Len. Toward a Cognitive Semantics. MIT Press 2000.