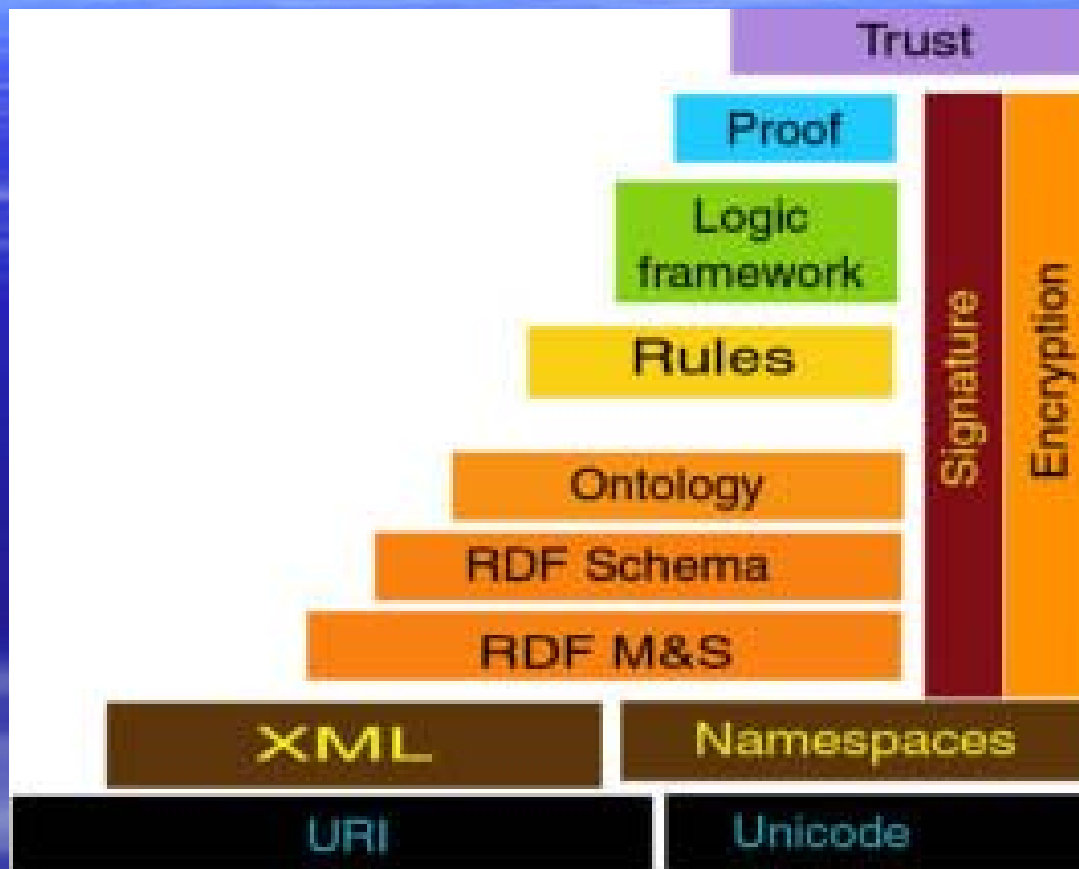


The Semantic Web: the apotheosis of linguistic annotation, or the basis of eScience?

Yorick Wilks
Oxford Internet Institute
and
University of Sheffield
www.dcs.shef.ac.uk

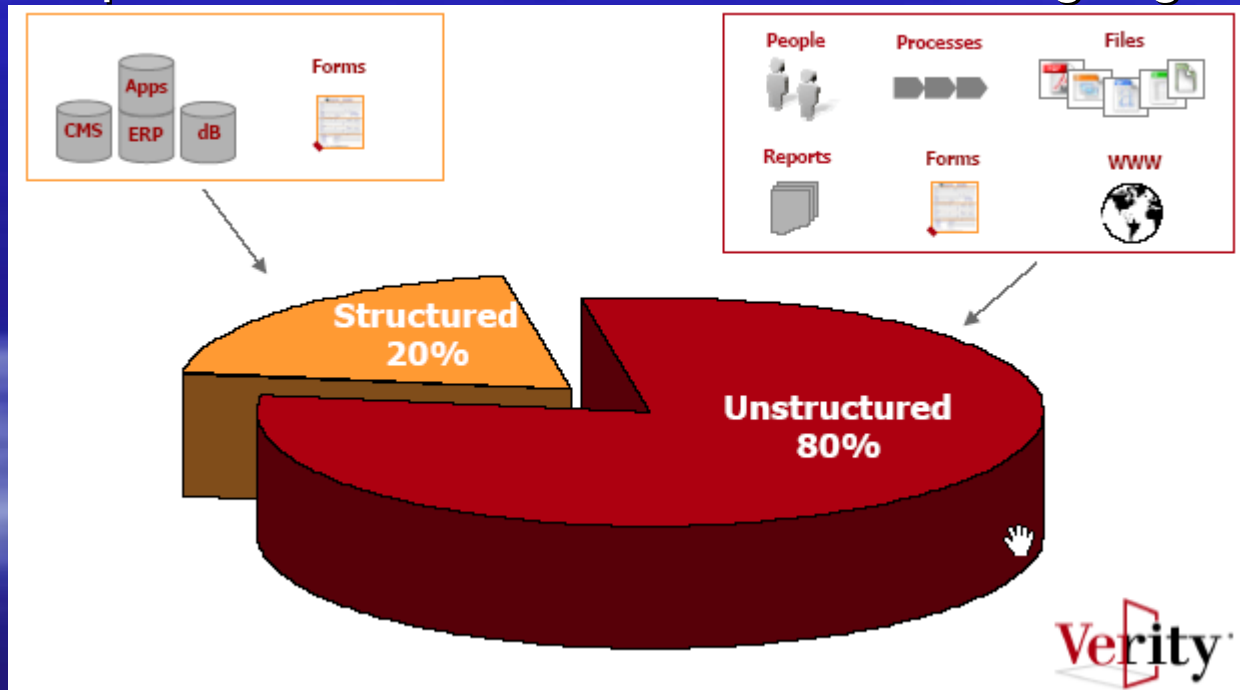


The Semantic Web and AI/NLP

- Some look at the top of the SW-pyramid and say “it’s just GOFAI isnt it”
- My case will be that if you look at the bottom of the pyramid, the SW rests much more on NLP than is usually realised.
- Of course, NLP people would say that, wouldn’t they, but they may be right!
- But here’s another reason for keeping NLP in mind.....

Sources of Knowledge

- 80-85% of a company's knowledge is contained in unstructured form,
 - i.e. expressed in some forms of natural language.



Talk topic: NLP and the Semantic Web

- NLP and annotation
- Automated IE as the engine at the bottom of the SW
- NLP and ontologies--the critique of the top level, and the empirical way outl.
- Examples from eBiology
- The whole web as a model for language
 - Do we need it?
 - Can we get it?

- “In the middle of a cloudy thing is another cloudy thing, and within that another cloudy and thing, inside which is yet another cloudy thing.....
-and in that is yet another cloudy thing, inside which is something perfectly clear and definite.”

-----Ancient Sufi saying.

Reminder: the sources of annotation:

- Roff, nroff, troff--publishing languages superceded by Knuth's TeX (about 1972): adding to a text ***how it should look***.
- The empirical NLP movement starts with POS tagging (CLAWS4, Leech, about 1979): ***adding to a text "what it means"***.
- The Text Encoding Initiative (1985?), then SGML, ***adding both to a text., but within the text***
- The metadata concept came from the DARPA NLP programs--annotations separate from the text.
- A subset of SGML later becomes HTML, then XML for the Web, then anything-at-all--ML (e.g. VoiceML).

Blurring the program and text distinction with XML.

- This distinction already blurred historically from both directions:
 - 1. Texts are really programs (one form of GOF AI)
 - 2. Programs are really texts
- And there's something in both these views

“Texts are really programs”:

- Hewitt’s wall-building program in Rustin’s NLP book 1972: “Natural Language is a side-effect”
- Longuet-Higgins (and others): “English is just a high level programming language” (1975?)
- Dijkstra: “ CS cannot really deal with natural language (i.e. as NLP) , because the latter isnt really up to its job” [YW’s version of ED’s views].

“Programs are really texts”:

- The “Wittgensteinian opposition” contains people like YW:
- “Understanding without proofs” IJCAI 1973
- “Programs and texts” 1977
- “What’s in a symbol” (with S. Nirenburg) JETAI 2000
- Bad fellow travellers (Derrida et hoc genus omnes)
- The parasitic view of programs and logic vs. NL (i.e. NL could exist without the others, and did for millennia, but not vice versa). On this view, logic/programs remain parasitic upon it in varying degrees, in that both retain language-like features.
- Look at NIL in LISP! [=FALSE =()]

Annotations as the most recent bridge from language to programs and logic

- The switcheroo is from text-infix annotations to meta-data considered separate from the text itself
- The former still corresponds to something linguistic and within NLP
- The latter conforms to old-AI --the McCarthy-Hayes program (1969)----and the dispensability of the text
- **This is the key notion:** Remember this the assumption of all true interlinguas (e.g. Schank), and is Sparck Jones' case against SW/AI/NLP

The Semantic web

(Berners-Lee, Hendler, and Lassila, Scientific American 2001)

- A vision of making the Internet as readable by computers (agents) as it is by us.
- A similar notion to the “ascent” in the semantic web pyramid---meaning/interpretation somehow tricking UP it from the bottom (cf. Braithwaite’s view of scientific theories--neutrinos linked to experiment)
- Is this last what SW people mean/want, or do they assume that the higher-level structures are self-interpreting?
- The SW as the basis of eScience?

Forms of knowledge in the SW

- 1) Universal Resource Indicators (URIs)
- 2) Resource Description Format
- RDF triples---putting all facts in the form:
 - John-LOVES-Mary
 - Not quite logic yet! Basically IE output; explicitly called « subject » « object »!
- 3) Ontologies---trees of concepts in hierarchical and functional relations: again like
 - Canary-ISA-Bird
- 4) DAML/OIL reasoning languages

The bottom levels of the SW always sit on NLP annotations of objects and actions (i.e. IE)

- Classic IE detects named entities--populates SW's "namespace"
- Does semantic type annotation
- Detects actions
- All recent IE works with an ontology
- It is also, of course, a SW annotator and RDF finder

Annotation Engines

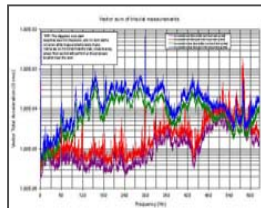
- Manual document annotation is still largely expected to be the main SW vehicle creation
 - Especially for trusted environments (e.g. within a company) this is expected to provide high quality material
- Automatic annotation is a vision but largely fulfilled:
 - To help manual annotation OR
 - To replace human annotators
- Producing automatic annotation services
 - For a specific ontological component/application
 - Constantly re-indexing documents

Melita (UShef)

- Document annotation assisted by adaptive IE
- Users:
 - Provides DAML ontology
 - Annotates document samples
- IE system:
 - Trains while users annotate
 - Generalizes over seen cases
 - Provides preliminary annotation for new documents
- Industrial users:
 - Solcara (UK): commercialisation+2 applications
 - Boeing (US)
 - Quinary (I)

X-MEDIA

Knowledge Sharing and Reuse across Media



European SW projects tend to accept the NL-based view of the SW

- E.g. X-Media (Ciravegna, Sheffield) a large EU 6FP IST 2005-9 on multi-media web services
- Based on IE/NLP annotation engines and machine learning.
- BUT, US and eScience projects still see the SW more in logic/AI based terms.

The philosophical problems may or may not just vanish as we push ahead with annotations!

- David Lewis and « markerese »: his 1970s critique of Fodor and Katz, and of any non-formal semantics (such as semantic type annotation)
- The Semantic Web takes this head on and carries on, hoping URIs and « popping out of the virtual world » (e.g. giving the web your phone number!) will solve semantic problems.
- Can all you want to know be put in RDF triples, and do the reasoning with them?
- But agents so-based do seem to work--*Eppur si Muove!*

An IR-based critique of the SW

- Sparck Jones, K. (2004) What's new about the Semantic Web? ACM SIGIR Forum.
- “words stand for themselves”: the basis of successful IR search in WWW and elsewhere.
- Content cannot be recoded in a general way especially if it is general content--IR has gained from “decreasing ontological expressiveness”
- Successful QA and IE are “superficial”
- Note: philosophical critics of types, like Lewis, want just OBJECTS, whereas Sparck Jones wants just WORDS.

The problem of Recoding content

- Charniak argued in 1974 that Schank's conceptual primitives such as [PTRANS liquid skin] could not distinguish sweat/sneeze/spit etc.
- (KSJ auction catalog example)"A Charles II parcel-gilt cagework cup, circa 1670"
- What, she asks, can be recoded beyond {object type: CUP}?
- The rest, she says, IS THE ENGLISH

Ameliorating the problem of grounding the meaning of SW terms (e.g. ontology terms)

- Inducing ontologies empirically from text
 - Can it be done?
 - Abraxas/Adaptiva
- Deriving very large language models to yield RDF-like “forms of fact” from text
- These are both special forms of IE (information extraction)

Abraxas/Adaptiva (EPSRC 04-07)

Christopher Brewster and Yorick Wilks

- Input:
 - A corpus
 - A seed ontology, or some pairs of terms
 - A relation chosen or labelled by the user
- Output:
 - A set of pairs of terms associated with labelled lexico-syntactic patterns
 - An extended ontology
- Key concept is an effective User Interface to allow user validation/ training of the system
- How partial/personal would such a system be?

Your very own Semantic Web?

Your very own Ontology?

- Former may be better than the latter
- Though the latter must be in our heads--we just cant get at them (thank goodness!)
- AI History of Points-of-View and To-hell-with-consistency
 - KRL
 - Viewgen
 - Hewitt's Programmar
 - De Kleer and Truth Maintenance
- Google-as-truth is also a POV phenomenon
- Scientific theories can have competing websites!
- Ontologies used to be personal and derived from just thinking!

All kinds of things and notions, to which names are to be assigned, may be distributed into such as are either more

{ *General*; namely those Universal notions, whether belonging more properly to

{ *Things*; called TRANSCENDENTAL } GENERAL. I
 { RELATION MIXED. II
 { RELATION OF ACTION. III

{ *Words*; DISCOURSE. IV

{ *Special*; denoting either

{ CREATOR. V

{ *Creature*; namely such things as were either *created* or *concreated* by God, not excluding several of those notions, which are framed by the minds of men, considered either

{ *Collectively*; WORLD. VI

{ *Distributively*; according to the several kinds of Beings. whether such as do

{ *Substance*; (belong to

{ *Inanimate*; ELEMENT. VII

{ *Animate*; considered according to their several

{ *Species*; whether

{ *Vegetative*

{ *Imperfect*; as *Minerals*, { STONE. VIII

{ METAL. IX

{ *Perfect*; as *Plant*, { HERB confid. accord. to the { LEAF. X

{ SHRUB. XIII { FLOWER. XI

{ TREE. XIV { SEED-VESSEL. XII

{ *Sensitive*; { EXANGUIOUS. XV

{ *Sanguineous*; { FISH. XVI

{ BIRD. XVII

{ *Parts*; { SPECULIAR. XIX { BEAST. XVIII

{ GENERAL. XX

{ *Accident*;

But having your own SW structure will come at a huge cost

- Negotiating meanings all the time by showing your ontologies and URIs
- These will have to stay pretty close or all communication will all fall apart---how much of conversation do you spend telling people what you mean by terms (if youre NOT a philosopher!)
- But it is the **personal data holdings** that companies now want access to and will organise for you!

Size of the Prize

The searchable Internet (in red) contains only a fraction as much information as the various other forms of digital media.

<i>Media type</i>	<i>Terabytes</i>	<i>Unique items per year</i>
Newsletters	1	40,000 titles
CD-ROMs	1	850 titles
Scholarly periodicals	6	37,609 titles
Books	39	950,000 titles
DVD videos	44	4,000 titles
Mass-market periodicals	52	80,000 titles
Audio CDs	58	33,443 titles
Newspapers	138	25,276 titles
Searchable Web	1.67	
Instant messaging	274	
Zip disks	350	1.4 million
Floppy disks	800	55 million
Office documents	1,397	10.75 billion pages
Audio MiniDisks	1,700	10.5 million
Flash memory	2,200	43 million
X-rays	20,000	2 billion
Motion pictures	25,000	10,342
Deep Web	91,850	
Audio tapes (analog)	128,800	128.8 million
Digital tapes	250,000	5 million
Photographs	375,000	75 billion
E-mails (originals)	440,606	
Digital video	1,265,000	1.15 million
Video tape (VHS and camcorder)	1,340,000	220 million
Hard disk drives	1,986,000	44 million

Taking stock here: three views of what the SW is:

- 1) An updating of the old AI dream of representing everything in logic for reasoning over the world (GOFAI); actually the SW is much less sophisticated than that--it has traded representation power for tractability.
- 2) An apotheosis of annotation in IE, attempting to build up to concepts in ontologies for e.g. scientific knowledge by very large shallow computations over texts: problem of grounding the terms other than in texts, and tying the general concepts plausibly to the distributions of usage in text.
- On this view the SW is the WW of text **plus meanings**.
- 3) A system of trusted data bases that ground meanings in something close to objects (TB-Ls own view?)
- This is close to Putnam's view that scientists are Guardians of Meaning

Putnam's scientists:

- Aluminium and Molybdenum look the same ----people cannot tell them apart but scientists can.
- Therefore scientists really know the meanings of these terms
- But they should not let this “leak out” to the general populace or the meanings might change.
- “cats from Mars” and the meaning of “cats”
- Compare: heavy water vs. water
- BUT (Hugh Mellor): we call heavy water “water” because it is water-----and that's simpler and more democratic than the alternative view
- You can't lock away meanings and keep them safe anywhere (as Wittgenstein might have said)--it's all usage in the end.

Is NLP experience any help at this point?

- Is the engineering end of the AI movement relevant here?? : empirical/corpus linguistics since 1990 and driven by:
 - speech research
 - Its extension to MT (Jelinek)
 - Information Retrieval (Sparck Jones)?
- “Taking words as they stand” (Sparck Jones); I.e. texts as “bags of words” with no
 - Logic
 - Linguistic features
 - Primitives
- But Jelinek’s MT adventure (only 50%) showed the limits of this-----was this lack of data or the need for empirically grounded structures?
- Jelinek has argued for both.

The importance of empirical/statistical studies of language distribution is that:

- Jelinek did get 50% of translations right without any “knowledge” of a foreign language---just trigram analysis.
- It does drive all the retrieval we have (from IR to Google)--Sparck Jones' point
- It may not give you concepts but it can, say, order translations by faithfulness, only with trigrams and without seeing the original.

A diversion from Wittgenstein to experiment:

- Wittgenstein said ask for the use not the meaning..
- Where would be better to look than the Web, as its usage is now so much larger than any human's language (60K years of reading!)
- Roger Moore on a hundred year's of human training to get a modern speech recognition system.
- Greffenstette's attempts to let web usage solve classic NLP problems:

Vague and distant relationship of Wittgenstein with classic AI:

- **"The solution to any problem in AI may be found in the writings of Wittgenstein, though the details of the implementation are sometimes rather sketchy."**
- Hirst, Graeme. "Context as a spurious concept." Proceedings, Conference on Intelligent Text Processing and Computational Linguistics, Mexico City, February 2000, 273--287.

Grefenstette's Experiments on MT and the Web

- Japanese to English

- アート : art, erte, rait, raitt, rat, rata, rate, ratte, ret, reta, rhett, rot, rott, rut, rutt
- クラブ : club, clubb, clubs, crab, crabb, crabbe, crabbs, crabs, craib, crave, craves, klebba, krabbe, krebbs, krebs, krob

- アートクラブ

- art club 135,287
- art clubs 4,825
- rot club 2,753
- rate club 2,371
- rat club 1,379
- rate clubs 317
- ret club 300
- rat clubs 227
- rott club 68
- rat crab 49

English
Web
Pages

Computation over the whole web- as-corpus can make usage a real concept:

- Greffentette's MT by voting
- BUT how much language to give a full model of a language?
- Language as a “system of very rare events” = data sparsity
- Trigram model may require a trillion words of training (=60,000 reading years)--loss of any connection to human language activity but still useful: see next figure-->



31/05/05

REVEAL

Lack of trigram coverage was above all what limited Jelinek's IBM MT project

- You can think about this as the way the repetitions of ngrams drop off with increasing n for a corpus of any imaginable size.
- A system that had noted COWS EAT and LIONS EAT would have no idea what to do with ELEPHANTS EAT (not to mention PRINTERS EAT PAPER).
- Jelinek himself became interested in what seem to be symbolic methods of classification to reduce this sparseness---e.g. semantic annotations and classifications.

But does the 1.5 billion word corpus (at 70%+ coverage) show things aren't so bad?

- By extrapolation it would need 75×10^{10} words to give 100% trigram coverage
- Our corpus at 74% was 15×10^8 , and Greffenstette calculated there were over 10^{11} words on English on the web in 2003 (i.e. about 12 times what Google now indexes).
- Since the whole web is hard to get at, could we go another way to improve coverage? Skipgrams?

Skip-Gram Example

Chelsea celebrate Premiership success.

tri-grams:

Chelsea celebrate Premiership
celebrate Premiership success

one-skip tri-grams:

Chelsea celebrate success

Chelsea Premiership success

The moral here is that recent work shows data sparsity for training may not be quite as bad as we thought:

- Trigram models give smooth natural output which rule grammars never can
- That was the main success of the IBM MT experiments
- Trigrams are very good models for speech, poor for meaning
- Skipgram models may compensate for size at the risk of nonsense--but that isn't so---see below:

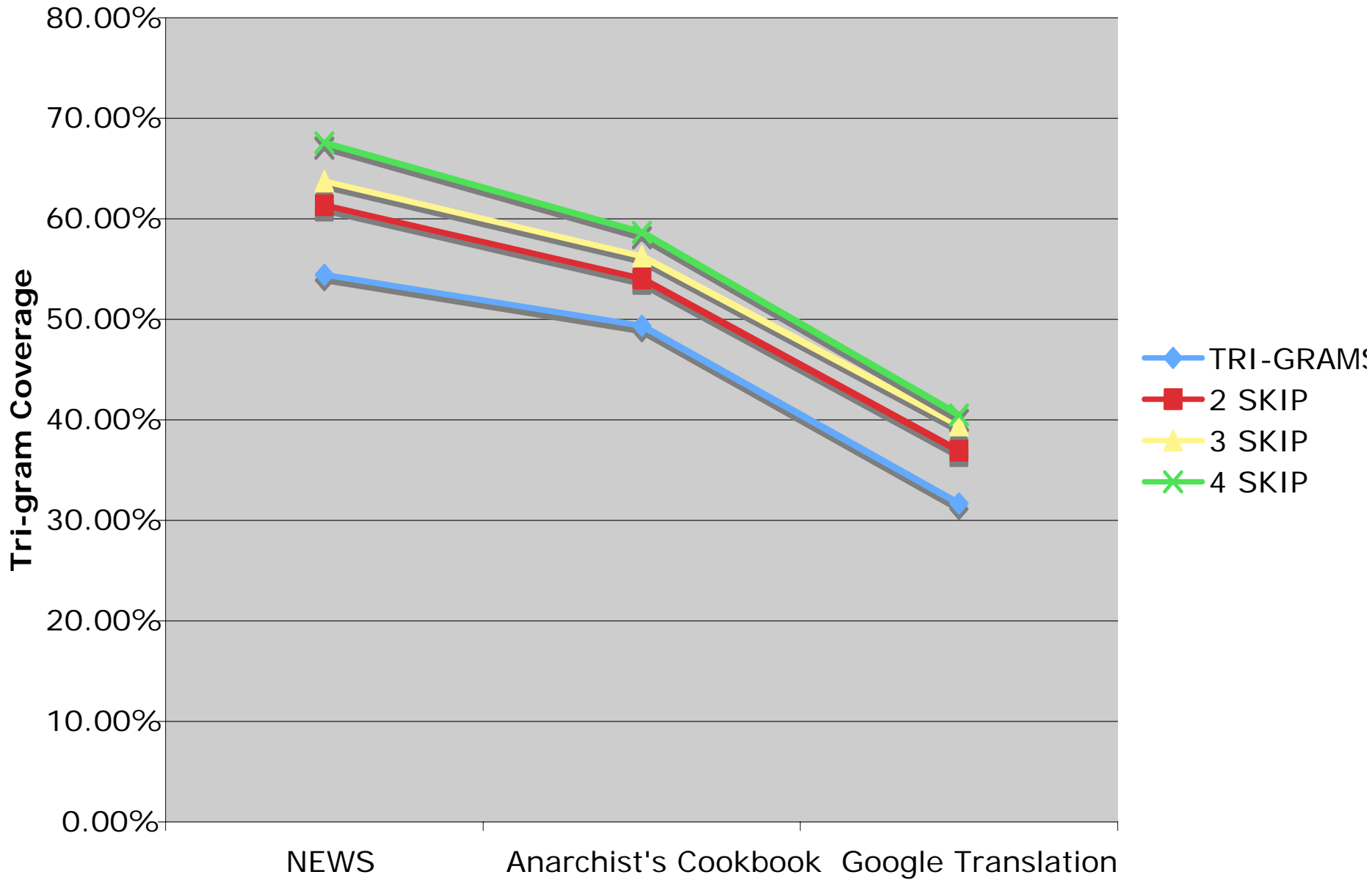
Previous Example

“*A bomb was deliberately detonated yesterday*” and

“*The bomb was detonated yesterday*”

- Using skip-grams, “**bomb detonated yesterday**” will be captured in both sentences.
- The number of trigrams captured within the text is increased, thus allowing a better model of context.

Skip-gram coverage on anomolous



Skip-grams prove their worth

- **Using skip-grams can be more effective than increasing the corpus size!**
- In the case of a 50 million word corpus, similar results are achieved using skip-grams as by quadrupling corpus size.
- This illustrates a serious possible use of skip-grams to expand contextual information

Preliminary conclusion on empirical results:

- Use empiricism may be stronger (with bigger corpora) than Jelinek thought in 1991
- But the corpora are so vast they cannot offer a model of how WE do semantics, so cognitive semantics remains an open question.
- Perhaps we can locate semantics within a world consisting only of word tokens, but which is not merely distributional?

So much for “meaning as (real, superhuman) use”--back into the cloudy stuff:

- All this what Sparck Jones would call doing NLP/CL by IR methods (AIJ, 2000)
- But what about those who want to go on talking about concepts and meanings.....?

What is the way out of wanting data/usage and concepts?

- Maybe we must take “words as the stand” (Sparck Jones) but perhaps not all words are equal
- Words as aristocrats, not democrats
- Perhaps “semantic primitives” are just words but also special words: forming a special language of translation, that is not pure but ambiguous, like all language.
- If that is so, perhaps we can have explanations, innateness (even definitions) on top of an empiricism of use.
- I would like this to be what “emergent semantics” is.

In that case:

- “primitive languages = languages of primitives” may or may not reflect the whole (game) language
- W. seems ambiguous as to whether (sub)parts of a language can represent the whole (see Nirenburg and Wilks, JETAI 2001).
- Can we smuggle back into a representation by empirical methods and “privileged words”:
 - Sophisticated information retrieval (time, space, colour, number....)
 - Some linguistic metadescription and innateness (over and above the learning mechanism that even empiricists allow)
 - Is this different from, say, the use of existing thesauri or Jelinek’s program to induce all linguistic structures??

More empirical semantics.

Imagine extracting from a vast corpus all forms of Agent-Action-Object triples (i.e. all examples of who does what to whom etc.).

- Use these to resolve ambiguity and interpretation problems of the kind that obsess people who are into concepts like ‘coercion’ ‘projection’, ‘metonymy’ etc. in lexical semantics.
- E.g. if in doubt what ‘my car drinks gasoline’ means, look at the stored triples about what cars do with gasoline and take a guess.

Forms-of-facts as a useful reality?

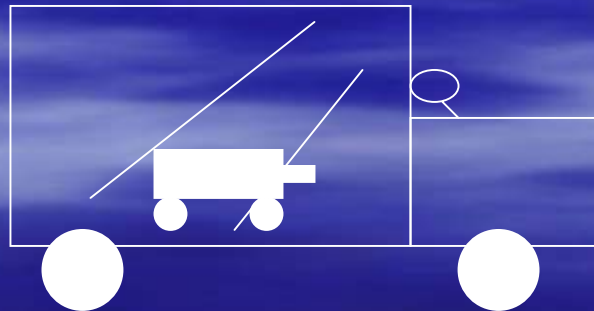
- This isn't a very good algorithm, but it should stir memories of Bar Hillel's (1959) argument against the very possibility of MT, namely that you couldn't store all the facts in the world you would need to interpret sentences
- AI has always believed you could
- Modern empirical corpus linguistics suggest you can find vast numbers of facts and use them.
- USC/ISI fact set currently around 2 million....
- Is this also a way of doing classic AI "knowledge-based understanding" on the cheap? CYC without People...!
- Is it any different from SW thinking based on RDF triples, except it delivers big numbers?

The man drove down the road in a car

((The man)(drove (down the road)(in a car))))



((The man)(drove(down the road(in a car))))



- For me this way of looking at things stirs quite different memories: of a more empirical version of the old Preference Semantics (1967) notion of doing interpretation by means of a list of all possible interlingual Agent-Action-Object triples! (only I made the list up!)
- These were intended then as Wittgensteinian forms-of-facts; but now they can be extracted automatically, by Information Extraction technology
- and are also very close to the RDF triples underlying the Semantic Web

What has any of this to do with eScience?

- T. Kazic (2006) Putting semantics into the Semantic Web: how well can it capture biology? Pacific Symposium on Biocomputing.
- “To sufficiently capture the scientific validity of the SW’s computations, it must sufficiently capture and use the semantics of the domain’s data and computations...it mustn’t confuse reactions of enzymes with reactions to drugs.

The data is extraordinary:

- “purine salvage” reactions
- $A \leftrightarrow B$
- $C \leftrightarrow D$
- An enzyme Z (EC 2.4.2.4) catalyses both but is not in class Y (purine nucleoside) and should not catalyse them (note in KEGG maps), as Z' (in Y) should, but it does.
- Moreover, Z promotes growth in some reactions and inhibits it in others (a statin).

The problem is how such data can be put in a SW ontology:

- GOFAI had “default logic” but this is more complex
- Particularly the opposite reactions in differing contexts.
- Can this be expressed other than by writing notes in the KEGG maps
- That is the challenge for the W in eScience
- Otherwise it's all LANGUAGE annotations in the “margins” of the structures which WE have to interpret (contra the SW hypothesis).

- Kazic: “..this implicit semantics is less effective than an explicit, by which I mean a computationally determinable, semantics”
- The same term in many URIs can only be disambiguated by definitions, but they are still in language.
- OR what “gene” means is implicitly programmed into the code manipulation gene descriptions; I.e. if NOT in language, then it cannot be understood or checked by scientists.
- Two ways out:
 - The SW learns to understand the language still present (the NL solution)
 - We give up worrying what the programs mean as long as they deliver (!)
 - GOFAI delivers after 40 years on a comprehensible, comprehensive, logical formalism.

Another case (FlyBase): the web for *Drosophila* studies

- FlyBase grounds gene identifiers (e.g. Rutabaga) in the genome, and associates an ID with e.g. regulatory regions, introns etc.
- BUT new regions are often identified for a gene “upstream” in the genome
- Thus the referent of the gene ID changes---and scientists live with this (I.e. what IS the gene ID of Rutabaga??)
- Cf. Putnam on guardians of meaning, and the shakiness of URIs in eScience. The moral is that there are not always formal URIs outside the SW, even in such areas of modern science.

Conclusion

- Contemporary NLP offers a way of looking at usage in detail and in quantity--even if those huge quantities now show we cannot easily relate them to an underlying theory of human learning and understanding.
- We can see glimmerings in machine learning studies of ‘language games’ in action, and of the role of key concepts in the representation of a whole language.
- Part of this will be some automated recapitulation of the role primitive concepts play in the organization of (human-built) ontologies, thesauri, wordnets.
- I argue NLP will continue to underlie the SW in several different ways: chiefly it is the only plausible way up to a defensible notion of meaning at conceptual levels,