Background and Foreground Knowledge in Dynamic Ontology Construction: Viewing Text as Knowledge Maintenance

Christopher Brewster, Fabio Ciravegna. and Yorick Wilks
NLP Group, Department of Computer Science
University of Sheffield, 211 Portebello Street
Sheffield, S1 4DP, United Kingdom
+44 114 2221967
{C.Brewster|F.Ciravegna|Y.Wilks}@dcs.shef.ac.uk

ABSTRACT

Ontologies have become a key component in the Semantic Web and Knowledge management. One accepted goal is to construct ontologies from a domain specific set of texts. An ontology reflects the background knowledge used in writing and reading a text. However, a text is an act of knowledge maintenance, in that it re-enforces the background assumptions, alters links and associations in the ontology, and adds new concepts. This means that background knowledge is rarely expressed in a machine interpretable manner. When it is, it is usually in the conceptual boundaries of the domain, e.g. in textbooks or when ideas are borrowed into other domains. We argue that a partial solution to this lies in searching external resources such as specialized glossaries and the internet. We show that a random selection of concept pairs from the Gene Ontology do not occur in a relevant corpus of texts from the journal Nature. In contrast, a significant proportion can be found on the internet. Thus, we conclude that sources external to the domain corpus are necessary for the automatic construction of ontologies.

1. INTRODUCTION

Ontologies are a key component both in the Semantic Web and in Knowledge Management. In the Semantic Web, they will provide a machine-interpretable knowledge infrastructure for a large variety of applications [2][6], including personal agents and B2B systems. In Knowledge Management, an ontology acts as a representation of an organisation's world view, as a 'corporate memory' and as a tool for the encoding of corporate experience and knowledge. While much has been written on their application and use, the real challenge lies in constructing them and keeping them up to date. Building ontologies is a complex and tedious process. It is labour intensive, error prone, and much like lexicography, as soon as the product is ready it is out of date. All this means that the cost is very high.

This paper is a contribution to efforts to automate or partially automate the ontology building process. Our starting point is that although ontologies attempt to represent the knowledge present in people's minds, the only easy access we have to what people think is through the texts they produce. A number of authors have been attempting to find ways to build ontologies from texts [22][15] [16]. The question we wish to raise here is to what extent is it a feasible enterprise, both *a priori* and in practice. There is a great deal of world knowledge which a reader brings to the interpretation of a text and we would like to try to determine the dividing line between what is explicitly expressed in a text (and thus could potentially be interpretable by a machine) and what is implicitly assumed.

This paper is organised as follows: Section 2 considers the relationship between texts and an ontology from a cognitive perspective. We argue that a text is normally re-enforcing and maintaining the knowledge assumed to be to be present in a reader's (hypothetical) ontology. In Section 3 we present an outline approach to overcome in part the absence of explicit interpretable context defining ontological knowledge. Section 4 presents some empirical data which indicate that such an approach may be fruitful; Section 5 considers related theoretical and empirical work, followed by a conclusion.

2. Ontologies and Text

There is not a great deal of agreement about ontologies other than that they are important. Ontologies vary widely in their complexity, formality and purpose. The most widely cited definition of an ontology is that of Gruber (1993) who said that an ontology was "a formal, explicit specification of a shared conceptualisation, used to help programs and humans share knowledge." The important aspect we wish to focus on here is that it is a *shared* set of concepts. For any given domain, the ontology is supposed to represent the concepts which are held in common by the participants in that domain. Thus it would appear that an ontology represents the background knowledge associated with a domain. It is this background knowledge to everyday life that *Cyc* (Lenat *et al.* 1994) attempts to encode, and most domain specific ontologies appear explicitly or implicitly to try to capture this knowledge. In a certain sense and without making any claims of psychological reality, there may be ontologies for each and every domain of human knowledge.

We may enquire then as to what the relationship is of a text to an ontology. It would be useful, in order to fulfil the needs of the

Semantic Web, Knowledge Management and a number of other application areas, to be able to construct automatically an ontology from a given set of texts. However, when a writer creates a text they assume a number of things. There is a linguistic assumption concerning the language used and a cognitive assumption concerning the ability of the audience to understand the vocabulary and technical terms used. In effect, a writer assumes that the audience shares the same or almost the same knowledge as themselves.

If the writer of a text assumes the same ontology as their own on the part of the audience, then the question arises as to what the purpose of an arbitrary text is. We would like to argue that one way of looking at a communicative act, a text, is to see it as an act of knowledge maintenance. There are three aspects to this:

- One aspect is that the text re-enforces the assumptions of background knowledge. It tells the reader which ontology to use to process the text (if we assume there exist different sub-ontologies) and re-enforces the knowledge structures of that ontology in the particular linguistic juxtaposition of concepts. For example, in the abstract quoted in Example 1, the use of the terms 'motor neuron', 'innervate', and 'transcription factors' immediately identify the domain and the respective background knowledge needed to understand the text.
- A second aspect is that the text alters the links, associations and instantiations of existing concepts. Thus a primary purpose of a text at some level is to change the relationship between existing concepts, or change the instantiations of those concepts. One way that texts provide 'new' information to the reader is by asserting a link previously absent, or by deleting a link previously assumed. This kind of activity can be seen as trying to re-structure the domain ontology which is clearly another form of knowledge maintenance. Again in Example 1, the phrase "*lim3* and *islet* constitute a combinatorial code that generates distinct motor-neuron identities" restructures the domain ontology.
- The third and most obvious way a text affects a domain ontology is by adding new concepts. The author may propose a new analytical concept or name a new procedure etc. and these acts of naming label new concepts and indicate their relationship with the rest of the ontology (as in Example 2)

Different classes of vertebrate motor neuron that innervate distinct muscle targets express unique combinations of LIM-homeodomain transcription factors[1, 2], suggesting that a combinatorial code of LIM-homeodomain proteins may underlie the control of motor-neuron pathway selection. Studies of LIM-homeodomain genes in mouse, *Drosophila melanogaster* and *Caenorhabditis elegans* have revealed functions of these genes in neuronal survival, axon guidance, neurotransmitter expression and neuronal function[3-8], but, ……..Our results provide evidence that *lim3* and *islet* constitute a combinatorial code that generates distinct motor-neuron identities.

Example 1, from Nature 397, 76 - 80 (1999)

A shell script is nothing more than a sequence of shell commands stuffed into a text file.

Example 2, from [21]

We would like to derive a complete ontology describing the knowledge of a domain from a set of texts. This would reach a holy grail in knowledge management, for example, since only texts allow some form of access into the minds of the writers and the knowledge stored there. However, the fact is that the background knowledge captured in an ontology is rarely explicitly stated in a text. It is implicit and taken for granted by the author. Consequently, it is very difficult to construct a computational processes which will capture what in essence is not there.

By explicit, we mean that a ontological relationship between two terms is expressed in some lexico-syntactic pattern of the type first identified by Hearst [11] over a decade ago. Examples include:

such NP as {NP_i, NP_2 {(or| and)} NP_n e.g.: …works by such authors as Herrick, Goldsmith, and Shakespeare.

NP, a NP that e.g. isolation and characterisation of pbp, a protein that interacts ….

NP and other NPs e.g. … malignant melanomas and other cancer cell types …

If we accept this line of thought, we should in principle find it extremely rare for background knowledge (i.e. the explicit specification of ontological relationships between two terms) to be explicitly expressed in any text. This we would expect to be especially true of a scientific text or academic paper because they are prototypical attempts to try to alter a community's accepted ontology. We would not expect this to be true of an introductory textbook, manual or glossary which by their nature do NOT assume the domain specific background knowledge (more accurately they assume a more general, possibly more top level, ontology).

It would follow that one would find a specification of an ontology at the borders of a domain. These borders might be in time or intellectual space. Thus, we might expect that when a concept or rather its corresponding lexicalisation is first introduced there will be statements defining or explicating the idea. On the other hand, when a concept is borrowed from one discipline into another, there again the term is likely to be defined.

3. Towards a Solution

There are two points here. Tthe foundation for efforts at automatically building ontologies from texts is the assumption that there are texts which do specify in a coherent manner the ontological relations one is interested in, and that these textual specifications can be read and processed by a machine.

Our hypothesis is that no matter how large our corpus, if it is domain specific, the major part of the domain ontology will not be specified because it is taken as given, or assumed to be part of

the background knowledge the reader brings to the text. This cannot be empirically proven because one could always imagine a larger collection of texts in a specific domain such that somewhere in it one might find the missing text which expresses the knowledge one is seeking identify. However, experience has shown that a certain number of textual contexts (citations) are needed for the ontological knowledge to be explicitly available. Thus, in view of Zipf's law, there will always be a tail end of terms of low frequency for which it is difficult to find sufficient or appropriate contexts within the corpus.

If we accept these two points, we can describe a model of text-based ontology construction as follows:

1. Initial input will be a set of texts for a specific domain
2. The system identifies which terms in the texts appear to be associated with which others. A number of methods exist for this including statistical, document distributional or using linguistic features [20] [4] [8].
3. The system attempts to identify ontologically relevant information from the text, in the first instance, using:
    a. Adaptive Information Extraction methods as described in [3] OR
    b. Manually identified lexico-syntactic patterns [7]
4. Where insufficient data is available in the text, or inappropriate data (i.e. sentential contexts which do not convey the ontological relations), the system identifies the need for further data
5. The system then searches textual sources external to the original domain specific corpus to overcome the absence of explicit communication of the background knowledge
6. Where no such context is found then either
    a. The potential ontological relation is rejected, OR
    b. The user of the system has to intervene

We are thus essentially constructing a model of ontology building where deficiencies of one data source (the original corpus) are expected to be compensated by other data sources. The key research problem then is identifying the 'correct' or 'appropriate' external source for a given set of texts.

There are a number of potential sources of such ontological knowledge, all of which present a certain challenges:

- **Encyclopaedias**: they might appear ideal sources for ontological knowledge. Clearly they include defining and explanatory texts which could be mined. The main problems are the difficulty of access to encyclopaedias and the fact that they are not likely to be very up to date. However, one can expect that the kind of background knowledge we are interested in does not change that rapidly.
- **Textbooks and manuals** associated with the domain also have potential usefulness. Here the main problem is identifying the relevant texts and obtaining them electronically. Furthermore, in both this case and that of the encyclopaedias, there is the problem of data sparseness – one will tend to find very few defining contexts.
- **Google Glossary**: this is a new experimental service from Google Labs which provides definition type texts for the terms one enters. For example:

Definitions for **Enzyme** from the web

- (n) 1. a protein which makes possible or facilitates a chemical reaction under a given environmental condition. 2. a digestive enzyme, an enzyme secreted by the body which helps break down food substances into forms that can be absorbed and assimilated by the body. Digestive enzymes are secreted by the salivary glands (e.g., amylase or ptyalin which breaks down starches); by the stomach (e.g. pepsin which breaks down proteins); by the liver (e.g., bile which help break down fats by emulsifyinng them); and by the pancreas (e.g., amylase which breaks down starches and lipase which breaks down fats.)
  http://prism.troyst.edu/~tiemeyep/glossary.htm
- Enzymes of the proteinic molecules occure in various reactions. They are biocatalysts, i.e. proteins allowing to increase the speed of a chemical reaction, at temperature compatible with the biological life (37 C). One of their properties is the specifity of action and reaction : each enzyme can be fixed only on one type of substrate (a molecule) and can catalyse only one chemical reaction. Once the catalysis is finished, the enzyme can enter in reaction again...
  http://library.thinkquest.org/26644/us/Lexique.htm
- Proteins that accelerate the rate of a specific chemical reaction without themselves being altered. Enzymes are generally named by adding the suffix "-ase" to the name of the substance on which the enzyme acts (for example, protease is an enzyme that acts on proteins).
  http://www.sahealthinfo.org/Modules/HIV_AIDS/aidsglossary/aidsglossary.htm

The main current problems with using Google Glossary as a source is that there are in a fact a great number of technical terms absent from it. Half of the 19 terms we consider below were absent from Google Glossary. Furthermore, we have no information as to how their lookup system works and it is unsatisfactory from a research perspective to use this type of black box.

- **the Internet**: this is the most obvious source and has both advantages and disadvantages:
    - **Advantages**

      a) It is extremely large so one is likely to find what is needed; b) It is continuously growing so recent conceptual developments are likely to be represented; c) It is easily accessed; d) We can understand what we can and cannot do with it
    - **Disadvantages**

      a) For any given term, texts can occur defining it in many different domains. Thus when looking for the genetic definition of 'chaperone', on the internet we find a number definitions (cf. Example 3). This is one form of noise; b) For narrow domains, even the internet does not cover the terminology – only the Deep Web.

a chaperone is a dangerous, freethinking individual in

Chaperone is a responsible female adult (minimum age of 21 years).

A chaperone is a helper protein that binds to a polypeptide …

Example 3

c) The perspective of a particular corpus may not correspond to that of the web as a whole, thus providing another form of noise. For example, 'metallochaperone ISA chaperone ISA molecular_function'according to the Gene Ontology, but looking for Hearst patterns we find that:

Within plant cells, chelating proteins such as metallochaperones allow the delivery of essential metal ions ….

Example 4

This tells us that 'metallochaperones' are proteins; d) The internet tends to repeat the same information in many places because people copy each other frequently, thus there is no guarantee that looking at yet another web site will give something new or useful; e) There is a major issue with respect to trust. It is difficult to determine criteria for deciding whether a web site is to be trusted or not. Also the same web site may be trustworthy for some types of information but untrustworthy for others.

There are specific parts of the internet which are more appropriate for our needs than others. For example, it may be possible to emulate Google in identifying glossaries, and attempt to match the glossary with the domain. For example, one would like to identify biology and medical glossaries for accessing when building an ontology like the Gene Ontology (www.geneontology.org). The main problems with using glossaries are the following: a) As hand built data structures they will not be very up to date (but as we noted with respect encyclopedias, this may not matter); b) It is hard to access and use many such glossaries as they often constitute part of the 'Deep Web' i.e. they are only accessible via some form or cgi script. This means wrappers may need to be constructed for each glossary; c) Glossaries have a peculiar perspective. They reflect the idiosyncracies of the creator(s) and often include or exclude terms arbitrarily.

4. Some Data

In order to provide initial evidence in favour of the hypothesis proposed in Section 3, we will show that it is impossible to reconstruct a 'gold standard' ontology like the Gene Ontology from a relevant collection of texts. We chose ten arbitrary pairs of terms, five from the 'higher' regions and five from the 'lower' regions, i.e. at or close to the leaves of the tree. No scientific claims are made for the manner in which the terms were chosen or the distinction between the higher and lower regions of the ontology since the 'depth' (i.e. the number of steps between a leaf and the root) is immensely variable. In order to simplify matters further, we limited ourselves to terms which were related by the IS-A relation.

We chose as our corpus all articles from the journal Nature covering a period from 1996 to 2001. Our domain specific corpus was a subset of these which concerned genetic topics (e.g. included the words *gene, genome, genetics,* etc.). Thus of 13000 texts in the whole corpus the subcorpus consisted of 1300 texts, or about 10%. We chose this ontology-corpus pair on the basis that it might be reasonable to use the Gene Ontology as a Gold Standard and attempt to determine how much of such an ontology could be derived *a priori* from a corpus like the journal Nature, which is the most prestigious journal in a range of bio-related fields.

We were looking for simple lexico-syntactic patterns of the type first suggested by Marti Hearst. If the corpus was theoretically to be a potential source of knowledge in order to construct an ontology, then at least some of the time ontological relationships would be expressed in the corpus by these sorts of patterns.

We observed the following frequency of distribution for our selected terms:

Table 1

Terms	Frequency	Common environments
histolysis isa	0	0
tissue death	0	
flocculation isa	0	0
cell communication	8	
vasoconstriction isa	0	0
circulation	42	
holin isa	0	0
autolysin	8	
aminopeptidase isa	3	0
peptidase	14	
death isa	654	0
biological process	12	
metallochaperone isa	0	0
chaperone	50	
hydrolase isa	9	2
enzyme	672	
ligase isa	92	2
enzyme	672	
conotoxin isa	1	0
neurotoxin	3	

It is clear from the above figures that there was no question of looking for lexico-syntactic environments since the terms hardly ever occurred in each other's environment.

So at this stage it appears reasonable to turn to other sources. In order to do this, we looked up each pair of terms in the following sources:

- Google (www.google.com)

- Google Glossary (http://labs.google.com/glossary)
- Encyclopaedia Britannica (www.britannica.com)
- Dictionary of Encyclopedia Britannica (www.britannica.com)

The results varied enormously depending on the terms. Some pairs were only to be found in online versions of the Gene Ontology, while at the other extreme over 31000 citations could be found on Google. In each case, the citations were checked manually for Hearst type patterns which could be said to explicitly represent the ontological relation between the terms. The individual results are presented in the Appendix, and the overall results are shown in Table 2.

Table 2

Textual Source	Number of contexts/articles found	Clear specification of ontological relation (no. of cases out of the 10 pairs of terms)
Original corpus	0 – 2	0
Google citations	3 - 31,000	6/10
Encyclopaedia	7/10	2/10
Dictionary	7/10	3/10

It is clear that using the internet directly provides the most likelihood of finding defining contexts. Using an encyclopaedia or dictionary appear to be no guarantee that definitions relating the two terms will be found. 60% of terms were found in explicit contexts on the internet which is clearly a great improvement on 0% in the original corpus. The figure of 60% does imply a limit on what is likely to be explicitly expressed, although a more systematic survey may give more reliable results. These figures should be taken merely as indicative.

Even though this is a limited sample, the figures appear to show that data cannot be derived from a corpus to demonstrate the expected ontological relationships. Even in the case of *ligase/enzyme,* where the number of occurrences in the text was respectable, the number of contexts where they co-occurred was very few. This could be seen as a simple data sparseness problem and the obvious solution is to increase the size of the corpus. But due to Zipf's law there will always be a very large proportion of the vocabulary in a given corpus which will occur too infrequently so as to provide opportunities for the knowledge concerning those terms to be explicitly stated.

We conclude that events where terms co-occur in a domain specific corpus are too sparse so as to provide sufficient opportunities for machine interpretable contexts to arise. Only by accessing external sources of information is there a significant increase in usable contexts.

5. Related Work

The term "background knowledge" is used loosely across a range of academic disciplines without receiving a precise definition. This is significant because knowledge can only be 'background' relative to some 'foreground'. "Background knowledge" is not to be confused with the term "tacit knowledge" which is widely used in knowledge management, and which refers to knowledge which is "semiconscious and unconscious knowledge held in peoples' heads and bodies" [14] as opposed to explicit knowledge which is "structured and accessible to people other than the individuals originating it" (ibid.). Hildreth and Kimble [12] present a critique of the tacit/explicit knowledge assumptions in knowledge management.

Some work has been undertaken analysing the process of knowledge maintenance mostly from the perspective of formal knowledge in an expert system or knowledge base [17][18] but this differs from the approach taken in this paper which views every text as an act of knowledge maintenance.

Morin [7][19] has built a system, PROMÉTHÉE, which attempts to learn the patterns Hearst [11] identified. His approach depended on repeated manual intervention unlike that presented in [3]. Maedche and Staab [15][16] present an ontology building environment which uses a number of algorithms and also external dictionaries to extract ontological knowledge from texts. However, apart from dictionaries they do not use external sources of information for the ontology learning process. Agirre *et al.* [1] use the internet to extend an existing ontology (WordNet) but they made no attempt to use the internet as a resource to specify the nature of the relationship between terms selected for inclusion. A key inspiration for the work presented here has been the Armadillo system developed by Ciravenga *et al.* [5], which uses external access to data available on the internet in order to complete the textual knowledge needed for an information extraction task.

6. Conclusion

This paper has argued that there is an inherent contradiction in the desire to build ontologies for a domain from a specific set of documents. There are cognitive reasons why this should be the case. The ontology reflects the assumed background knowledge which the text is 'maintaining' i.e. re-enforcing and modifying. Furthermore, there is the practical reality that terms do not co-occur sufficiently frequently so as to make possible the machine interpretation of the requisite knowledge. The partial solution proposed in this paper is to use external sources such as the internet, and possibly more domain relevant sources, to compensate for the knowledge gaps in the initial corpus. We have argued that, in one limited experiment at least, we can go from no exemplars to at least 60% of cases having exemplars for the relevant knowledge relations.

There are various potential responses. One approach would be to take this as clear demarcation of the limits of what computational approaches can derive from text. Hays [10] has argued, with respect to disambiguation, that there are limits as to what is linguistically encoded in a text and the rest is dependant on the world knowledge the reader brings. However, another perspective is that the limits of parsing and processing texts have yet to be reached so it is far too early to say what is and is not interpretable by machine.

Future work will involve determining which lexico-syntactic patterns are most appropriate for exploring a particular ontological relation and possibly developing a voting mechanism. Specialised resources need to be identified and while this can be done manually in a specific case, it would be beneficial to develop automated methods for identifying appropriate external resources. Finally, we need to deal with conflicting responses from different resources. There needs to be found ways to evaluate a resource so as to measure its trustworthiness and to resolve conflicting responses from different resources.

7. ACKNOWLEDGMENTS

This work is supported under the Advanced Knowledge Technologies (AKT) Interdisciplinary Research Collaboration (IRC), which is sponsored by the UK Engineering and Physical Sciences Research Council under grant number GR/N15764/01. The AKT IRC comprises the Universities of Aberdeen, Edinburgh, Sheffield, Southampton and The Open University.

8. REFERENCES

[1] Agirre, E., O. Ansa, E. Hovy, & D. Martínez. Enriching very large ontologies using the WWW, Proceedings of the ECAI 2000 workshop "Ontology Learning" 2000

[2] Berners-Lee, T., J. Hendler, O. Lassila. The Semantic Web, in Scientific American, Issue 501 (2001) http://www.sciam.com/2001/0501issue/0501berners-lee.html

[3] Brewster, C., F. Ciravegna and Y. Wilks User-Centred Ontology Learning for Knowledge Management *7th International Workshop on Applications of Natural Language to Information Systems*, Stockholm, June 27-28, 2002, Lecture Notes in Computer Sciences, Springer Verlag.

[4] Brown, P. F., V. J. Della Pietra, P. V. DeSouza, J. C. Lai, & R. L. Mercer. Class-based n-gram models of natural language, Computational Linguistics, 18 (1992), 467-479

[5] Ciravegna, F., A. Dingli, D. Guthrie, and Y. Wilks. Mining Web Sites Using Unsupervised Adaptive Information Extraction. Proceedings of the 10th EACL-03 (Budapest, Hungary, April 2003)

[6] Fensel, D., F. van Harmelen, I. Horrocks, D.L. McGuiness, & P.F. Patel-Schneider. OIL: An Ontology Infrastructure for the Semantic Web, IEEE Intelligent Systems (16) 2001

[7] Finkelstein-Landau, M., & E. Morin. Extracting Semantic Relationships between Terms: Supervised vs. Unsupervised Methods, Proc. International Workshop on Ontological Engineering on the Global Information Infrastructure, 71-80, Dagstuhl Castle, Germany, (1999).

[8] Grefenstette, G. Explorations in Automatic Thesaurus Discovery, Amsterdam: Kluwer, 1994

[9] Gruber, T. A translation approach to portable ontology specification, Knowledge Acquisition 5 (1993), 199-220

[10] Hays, Paul R. Collocational Similarity: Emergent Patterns in Lexical Environments, Dissertation submitted to the School of English, University of Birmingham, 1997

[11] Hearst, M.A. Automatic Acquisition of Hyponyms from Large Text Corpora, in *Proceedings of COLING 92*, Nantes, 1992

[12] Hildreth, P.M. & C. Kimble. The Duality of Knowledge, in Information Research, Vol. 8 No. 1, October 2002

[13] Lenat, D. B., R. V. Guha, K. Pittman, D. Pratt, & M. Shepherd. Cyc: Toward Programs With Common Sense, (1994) available at http://www.cyc.com/tech-reports/act-cyc-108-90/act-cyc-108-90.html

[14] Leonard, D. & Sensiper, S.The role of tacit knowledge in group innovation. California Management Review, **40**(3), 112-132 (1998)

[15] Maedche, A. and S. Staab. Mining Ontologies from Text, in Rose Dieng, Olivier Corby (Eds.): Knowledge Acquisition, Modeling and Management, 12th International Conference, EKAW 2000, Juan-les-Pins, France, October 2-6, 2000, Proceedings. Lecture Notes in Computer Science 1937 Springer

[16] Maedche, A. and S. Staab. Learning Ontologies for the Semantic Web, in Stefan Decker, Dieter Fensel, Amit P. Sheth, Steffen Staab (Eds.): *Proceedings of the Second International Workshop on the Semantic Web - SemWeb'2001*, Hongkong, China, May 1, 2001. http://CEUR-WS.org/Vol-40/

[17] Menzies, T.J. Knowledge Maintenance: The State of the Art, *The Knowledge Engineering* Review, 10(2), 1998. Available from http:// menzies.us/pdf/97kmall.pdf

[18] Menzies, T.J. and J. Debenham, Expert Systems Maintenance; *Encyclopedia of Computer Science and Technology*; pages 35-54; volume 47; number 27; Marcell Dekker Inc.(2000); Available from http://menzies.us/pdf/00cst.pdf

[19] Morin, E. Using Lexico-Syntactic patterns to Extract Semantic Relations between Terms from Technical Corpus, *TKE 99*, 268-278, Innsbruck, Austria, (1999)

[20] Scott, M. Focusing on the Text and Its Key Words, *TALC 98 Proceedings*, (1998) Oxford, Humanities Computing Unit, Oxford University.

[21] Schwartz, R. and T. Christiansen, *Learning Perl*, O'Reilly: Sebastopol, CA, 1997

[22] Wilks, Y. and S. Nirenburg. Towards Automated Knowledge Acquistion. KB&KS '93 (December 1993) Tokyo, Japan

9. Appendix

Terms:	**histolysis** isa **tissue death**
Textual Source	
Original corpus	0
Google citations	3, all from the 'Gene Ontology'
Google glossary	0

Encyclopaedia (Britannica)	1, under 'lepidopteran' "tissues of the larva undergo considerable **histolysis** (breaking down)"
Dictionary (Britannica)	histolysis: "the breakdown of bodily tissues"

Terms:	**flocculation** isa **cell communication**
Textual Source	
Original corpus	
Google citations	31, of which about 10 from 'GO', but no defining contexts
Google glossary	9
Encyclopaedia (Britannica)	7 references none concerning cell communication
Dictionary (Britannica)	flocculate: to cause to aggregate into a flocculent mass :to become flocculent

Terms:	**vasoconstriction** isa **circulation** (nb. the GO appears to be clearly wrong here)
Textual Source	
Original corpus	0
Google citations	17k, many examples showing a close relationship but NOT the one specified in the ontology
Google glossary	9
Encyclopaedia (Britannica)	11, Sub-article on vasoconstriction which implies it is a disease of the arteries,
Dictionary (Britannica)	vasoconstriction: **:** narrowing of the lumen of blood vessels especially as a result of vasomotor action

Terms:	**holin** isa **autolysin**
Textual Source	
Original corpus	0
Google citations	30, including GO references, showing a close association but no ontologically clear relationship. Citations show "holing is a protein"
Google glossary	0
Encyclopaedia (Britannica)	0
Dictionary (Britannica)	0

Terms:	**aminopeptidase** isa **peptidase**
Textual Source	
Original corpus	0
Google citations	7k, which tell us aminopeptidase is an enzyme not that it is a peptidase
Google glossary	0
Encyclopaedia (Britannica)	1, which can humanly be understood to convey that aminopeptidase is an enzyme
Dictionary (Britannica)	aminopeptidase: **:** an enzyme that hydrolyzes peptides by acting on the peptide bond next to a terminal amino acid containing a free amino group

Terms:	**death** isa **biological process**
Textual Source	
Original corpus	0
Google citations	6k, including 4 specifying that "death is a biological process", "biological processes"/death 350k, many for "biological processes such as cell death"
Google glossary	8
Encyclopaedia (Britannica)	7, none helpful, however in article on 'death', it is humanly understandable that 'death is a biological process'
Dictionary (Britannica)	death: **:** a permanent cessation of all vital functions **:** the end of life

Terms:	**metallochaperone** isa **chaperone**
Textual Source	
Original corpus	0
Google citations	257, including many references to GO, but none clearly specifying this ontological relationship
Google glossary	0
Encyclopaedia (Britannica)	0 (no 'metallochaperone', and no biological reference to 'chaperone')
Dictionary (Britannica)	

Terms:	**hydrolase** isa **enzyme**

Textual Source	
Original corpus	2
Google citations	29k, many contexts where this is derivable: "hydrolase is a ubiquitous cellular enzyme" but also: "hydrolase is a protease" "hydrolase is a peroxisomal coenzyme"
Google glossary	0
Encyclopaedia (Britannica)	1 article, with definition: "any one of a class of more than 200 enzymes that …"
Dictionary (Britannica)	hydrolase: **:** a hydrolytic enzyme

Terms:	**ligase** isa **enzyme**

Textual Source	
Original corpus	2
Google citations	31k, many contexts where this is derivable: "DNA ligase: Enzyme involved in the replication and repair" but also: "ligase is a single polypeptide" "ligase is a 600 kDa multisubunit protein"
Google glossary	9
Encyclopaedia (Britannica)	1 article with definition: "*also called* ***Synthetase*** any one of a class of about 50 enzymes that …."
Dictionary (Britannica)	ligase: "an enzyme that catalyzes the linking together of two molecules"