Automatic Extraction of Semantic Networks from Text using Leximancer

Andrew E. Smith.
Key Centre for Human Factors and Applied Cognitive Psychology,
The University of Queensland,
Queensland, Australia, 4072.
asmith@humanfactors.uq.edu.au

Abstract

Leximancer is a software system for performing conceptual analysis of text data in a largely language independent manner. The system is modelled on Content Analysis and provides unsupervised and supervised analysis using seeded concept classifiers. Unsupervised ontology discovery is a key component.

1 Method

The strategy used for conceptual mapping of text involves abstracting families of words to thesaurus concepts. These concepts are then used to classify text at a resolution of several sentences. The resulting concept tags are indexed to provide a document exploration environment for the user. A smaller number of simple concepts can index many more complex relationships by recording co-occurrences, and complex systems approaches can be applied to these systems of agents.

To achieve this, several novel algorithms were developed: a learning optimiser for automatically selecting, learning, and adapting a concept from the word usage within the text, and an asymmetric scaling process for generating a cluster map of concepts based on cooccurrence in the text.

Extensive evaluation has been performed on real document collections in collaboration with domain experts. The method adopted has been to perform parallel analyses with these experts and compare the results.

An outline of the algorithms (Smith, 2000) follows:

1. Text preparation: Standard techniques are employed, including name and term preservation, tokenisation, and the application of a stop-list.

2. Unsupervised and supervised ontology discovery: Concepts can be seeded by a domain expert to suit user requirements, or they can be chosen automatically using a ranking algorithm for finding seed words which reflect the themes present in the data. This process looks for words near the centre of local maxima in the lexical co-occurrence network.

3. Filling the thesaurus: A machine learning algorithm is used to find the relevant thesaurus words from the text data. This iterative optimiser, derived from a word disambiguation technique (Yarowsky, 1995), finds the nearest local maximum in the lexical cooccurrence network from each concept seed. Early results show that this lexical network can be reduced to a Scale-free and Small-world network[1].

4. Classification: Text is tagged with multiple concepts using the thesaurus, to a sentence resolution.

5. Mapping: The concepts and their relative cooccurrence frequencies now form a semantic network. This is scaled using an asymmetric scaling algorithm, and made into a lattice by ranking concepts by their connectedness, or centrality.

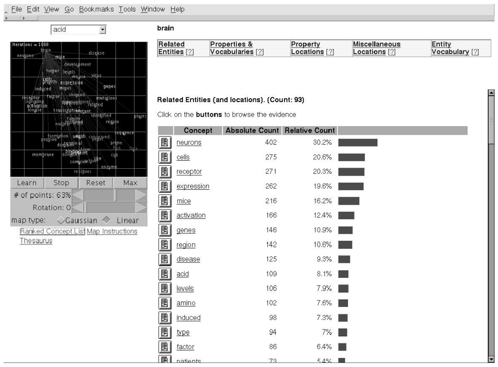
6. User interface: A browser is used for exploring the classification system in depth. The semantic lattice browser enables semantic characterisation of the data and discovery of indirect association. Concept co-occurrence spectra and themed text segment browsing are also provided.

2 Analysis of the PNAS Data Set

The data set presented here consisted of text and metadata from Proceedings of the National Academy of Science, 1997 to 2002. These examples are extracted from the abstract data. Firstly, Leximancer was configured to map the document set in unsupervised mode. A screen image of this interactive map is shown in figure 1. This

[1] Following (Steyvers and Tenenbaum, 2003).

shows the semantic lattice (left), with the co-occurrence links from the concept ‘brain’ highlighted (left and right).



brain

Related Entities [?]	Properties & Vocabularies [?]	Property Locations [?]	Miscellaneous Locations [?]	Entity Vocabulary [?]

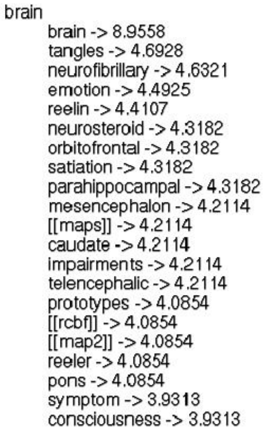
Related Entities (and locations). (Count: 93)

Click on the **buttons** to browse the evidence

	Concept	Absolute Count	Relative Count	
	neurons	402	30.2%	
	cells	275	20.6%	
	receptor	271	20.3%	
	expression	262	19.6%	
	mice	216	16.2%	
	activation	166	12.4%	
	genes	146	10.9%	
	region	142	10.6%	
	disease	125	9.3%	
	acid	109	8.1%	
	levels	106	7.9%	
	amino	102	7.6%	
	induced	98	7.3%	
	type	94	7%	
	factor	86	6.4%	
	patients	73	5.4%	

Figure 1: Unsupervised map of PNAS abstracts.

Figure 2 shows the top of the thesaurus entry for the concept ‘brain’. This concept was seeded with just the word ‘brain’ and then the learning system found a larger family of words and names which are strongly relevant to ‘brain’ in the these abstracts. In the figure, terms in square brackets are identified proper names, and numerical values are the relevancy weights.



brain
brain -> 8.9558
tangles -> 4.6928
neurofibrillary -> 4.6321
emotion -> 4.4925
reelin -> 4.4107
neurosteroid -> 4.3182
orbitofrontal -> 4.3182
satiation -> 4.3182
parahippocampal -> 4.3182
mesencephalon -> 4.2114
[[maps]] -> 4.2114
caudate -> 4.2114
impairments -> 4.2114
telencephalic -> 4.2114
prototypes -> 4.0854
[[rcbf]] -> 4.0854
[[map2]] -> 4.0854
reeler -> 4.0854
pons -> 4.0854
symptom -> 3.9313
consciousness -> 3.9313

Figure 2: Thesaurus entry for ‘brain’ (excerpt).

It is also of interest to discover which concepts tend to be unique to each year of the PNAS proceedings, and so identify trends. This usually requires a different form of analysis, since concepts which characterise the whole data set may not be good for discriminating parts. By placing the data for each year in a folder, Leximancer can tag each text sentence with the relevant year, and place each year as a *prior* concept on the map. The resulting map contains the prior concepts plus other concepts which are relevant to at least one of the priors, and shows trending from early years to later years (figure 3).

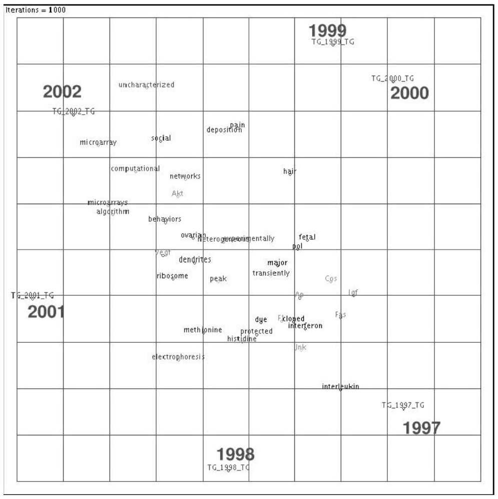


Figure 3: Temporal map of PNAS abstracts.

3 Conclusion

The Leximancer system has demonstrated several major strengths for text data analysis:

- Large amounts of text can be analysed rapidly in a quantitative manner. Text is quickly re-classified using different ontologies when needs change.
- The unsupervised analysis generates concepts which are well-defined — they have signifiers which communicate the meaning of each concept to the user.
- Machine Learning removes much of the need to revise thesauri as the domain vocabulary evolves.

References

Andrew E. Smith. 2000. Machine mapping of document collections: the leximancer system. In *Proceedings of the Fifth Australasian Document Computing Symposium*, Sunshine Coast, Australia, December. DSTC. http://www.leximancer.com/technology.html.

Mark Steyvers and Joshua B. Tenenbaum. 2003. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Submitted to Cognitive Science*. http://www-psych.stanford.edu/˜msteyver.

David Yarowsky. 1995. Unsupervised word-sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL-95)*, pages 189–196, Cambridge, MA. http://www.cs.jhu.edu/˜yarowsky/pubs.html.