The ACAMRIT[1] semantic tagging system: progress report.

Paul Rayson, Andrew Wilson.
UCREL,
Departments of Computing and Linguistics,
Lancaster University.

Abstract

Building on a successful previous project, UCREL (the University Centre for Computer Corpus Research on Language) is collaborating with Reflexions Communication Research (a market research company in London, UK) to develop software which will undertake the semantic tagging of words in a text, facilitate the assignment of 'content tags' to those words, and provide a statistical analysis of the resulting tag frequency profile. The project intends to extend previous work by developing enhanced disambiguation techniques, larger lexical resources and word sense frequency data for spoken English, automatic pronoun resolution and a broader dependency-style syntactic analysis.

This project aims to combine the best of both qualitative and quantitative survey research, but it has many other potential applications in linguistics and more generally in the social sciences and humanities.

Introduction

The ACAMRIT system is being built within the ACASD (Automatic Content Analysis of Spoken Discourse) project[2], which is a follow-up to a project with the same collaborators (see Wilson and Rayson, 1993).

ACAMRIT is a suite of programs for the automated semantic tagging and content analysis of spontaneous spoken English. The system is made up of several separate software modules. The main constituents are a formatting pre-processor (TOSGML), the CLAWS probabilistic part-of-speech tagging program previously developed at UCREL (see Garside, 1996), a semantic tagger (SEMTAG), a syntactico-semantic linking system (MATRIX), a pronoun resolution module (GENESIS), the statistics and retrieval package with a user-friendly X-windows interface (SEMSTAT), a program for manual postediting of the output (SEMEDIT), and a program for mapping semantic tags onto research-specific content categories (MAPPING). All the programs are written in C and designed to run on a Sun Workstation. Results from the retrieval package can be printed out, but the main feature for the end-user is the interactive nature of the retrieval process.

The system is being used by the market research company for commercial projects, so our aim is to make the system as robust as possible, i.e. so that any text one cares to

[1] Otherwise known as "Automatic Content Analysis of Market Research Interview Transcripts".
[2] The project is funded by the EPSRC (Engineering and Physical Sciences Research Council), and is part of a collaborative (JFIT) project with commercial partnership and the support of the DTI.

analyse would be processed. For this reason, our transcription guidelines are limited to producing raw text in normal orthographic form with a minimum of mark-up.

The software is used as an aid in preparing reports for market research clients based on a set of one-to-one non-directed interviews with members of the public selected for a given project. The results obtained are more flexible than the normal tick-the-box interviews conducted to produce quantitative estimates about a particular product or service: the interviewee is not limited to answering a series of questions set in advance. The results are also an improvement on a qualitative survey which consists of a small number of non-directed interviews: here, the market researcher cannot hand-analyse enough text in a given limited time period to make quantitative judgements about the product. We think our software gives the best of both worlds: quantitative estimates and the option of viewing the underlying text of the interviews.

Once the text of the interviews has been processed to the SEMTAG stage, we can produce frequency profiles of semantic tags which highlight statistically significant items for further investigation, perhaps by concordance. The χ^2 (Chi-squared) test is used to give a value to the words or tag frequencies which differ from a corpus-based norm value. We have collected nearly 3 million words of spoken data from people all over the country to produce these norms.

The normative corpora

Corpus A in total will be around 1 million words. It is still being collected and transcribed by Reflexions. Interviews will be conducted with people from 100 types of institutions. The set of institutions are designed to represent a cross-spectrum of British daily life within a broad set of topic areas:

Agriculture and Horticulture
Amenities
Architecture and Buildings
The Arts and Heritage
The Body and Health
Charity
Clothing and Fashion
Clubs and Societies
Communications
Constitution, Government and Politics
Education
The Family
Food and Drink
Games
Labour Relations
Law and Order
Media
The Military
Money and Business
Places of Residence
Recreation
Religion
Science and Technology
Sport
Travel and Transport

The aim of corpus A is to extend the vocabulary on which we train the system.

Corpus B was collected by Reflexions from 13 geographical regions of the UK (based on TV regions). 2 million words resulted from 797 interviewees, collected between January and June 1994. The interviews were non-directed, usually starting with the question "What's on your mind?". The idea was to collect a representative norm of English usage in a situations similar to that in which the product interviews are to be conducted. The people selected for interview form a balanced sample of age, gender, region, and social class in a similar way to the demographic part of the British National Corpus (BNC)[3]. Corpus B is partly marked up for anaphoric reference.

Raw input to the system

Here is an example of the transcription from an interview conducted in Birmingham.

```
<person age=59 sex=f type=C1 Region=Birmingham Interviewer=Judy
Origin=Midlands/Central>
<q>  Can we start off by just talking about what's on your mind?
<r>  Now?
<q>  Anything you want to talk about.
<r>  Well my holidays, yes, and what I did this week and whether it's
going to pour down with rain, it looks like it.
```

Each interview is headed by an SGML (Standard Generalised Mark-up Language) tag (the text inside angled brackets). We normally record age, gender, social class, region, the interviewer's name and a rough idea of the origin of the interviewee. In fact any attribute can be recorded in the same format. The header information is used by the retrieval software to split the data under analysis and produce sub-corpora if required. We mark questions and answers so that the interviewers' language can be omitted from the analysis but included in any concordances produced.

Batch processing

Each text is processed by a script (called 'sausage') which passes it through the various automatic modules we have:

1. TOSGML: checks the raw format and converts non-ASCII characters (such as accented letters) into a standard format.
2. CLAWS: (Constituent Likelihood Automatic Word-tagging System) developed (outside ACAMRIT) by Roger Garside and other members of UCREL at Lancaster University, using a combination of hidden Markov modelling and grammatical templates to assign part-of-speech (POS) tags to English texts.
3. NPEXT: developed by Paul Rayson (in an EU sponsored project ET10/63, see Garside et al 1993) to mark noun phrases automatically in POS tagged text. The noun phrase information is used to aid semantic idiom tagging and MATRIX.
4. SEMTAG: semantic tagger (see below).
5. GENESIS: currently undergoing trials to mark antecedents of 3rd person pronouns. The antecedents will be used in SEMSTAT to augment frequencies, mainly for product names.

[3] The BNC was compiled in the years 1991-1995 by a collaborative team consisting of Oxford University Press (the lead partner), Longman Group UK Ltd., Chambers Harrap, the British Library, and the Universities of Oxford (Oxford University Computing Services) and Lancaster (UCREL). For details of availability for research purposes please contact Oxford University Computing Services, 13 Banbury Road, Oxford OXN 6NN, UK. or email natcorp@oucs.ox.ac.uk. Major funding for the project was provided by the Science and Engineering Research Council and by the Department of Trade and Industry. For a description and rationale of the spoken component of the BNC, and in particular of the demographic corpus, see Crowdy 1993, and 1995.

6. MATRIX: assigns links between adjectives and the nouns they modify, degree modifiers and adjectives, and deals with negation including transferred negation (see Wilson 1991 and 1993). The links will be used by SEMSTAT to provide a high level of detail in the frequency analysis.

CLAWS tagging is roughly 96-98% correct on written texts without manual postediting; however, we do notice an increase in errors with spoken discourse. The noun phrase identification has an error rate of 15%, and the semantic tagger has an error rate of about 10%. The linking program is successful more than 90% of the time.

Semantic tagging

The semantic tagset was originally loosely based on Tom McArthur's Longman Lexicon of Contemporary English (McArthur, 1981). It has a multi-tier structure with 21 major discourse fields, subdivided, and with the possibility of further fine-grained subdivision in certain cases, for example:

```
E - EMOTIONAL ACTIONS, STATES AND PROCESSES
1 General
2 Liking
3 Calm/Violent/Angry
4 Happy/sad: 1 Happy
             2 Contentment
5 Fear/bravery/shock
6 Worry, concern, confident

H - ARCHITECTURE, BUILDINGS, HOUSES AND THE HOME
1 Architecture and kinds of houses and buildings
2 Parts of buildings
3 Areas around or near houses
4 Residence
5 Furniture and household fittings

L - LIFE AND LIVING THINGS
1 Life and living things
2 Living creatures generally
3 Plants
```

Antonyms are identified using +/- markers. So, for example, happy would usually be tagged E4.1+ and sad would be tagged E4.1-.

We have a lexicon of over 36,000 grammatical words (i.e. word and part-of-speech), and an idiom list with over 15,000 entries. The idioms are phrases like *all in all*, *art nouveau*, and *have a screw loose*, to which we assign a single semantic tag.

We apply the following set of disambiguation techniques to promote the correct tag to the head of the tag list on each word (as this is the one selected by SEMSTAT):

1. POS tag

Consider the word 'spring'. It could be a singular common noun 'NN1' (as in Slinky or a water source), a temporal noun 'NNT1' (i.e. the season), or a base form of lexical verb 'VV0' (as in spring upwards). As a singular common noun we would give it a general object 'O2' or geographical 'W3' semantic tag. As a temporal noun it would be tagged as 'T1.3' (time period), and as a verb it would be given the main sense 'M1' which is the semantic tag for moving, coming and going, and other tags (as below). The entry for spring in our lexicon is therefore:

```
spring                   NN1     O2 W3
spring                   NNT1    T1.3
spring                   VV0     M1 A2.2 Q2.1 G2.1-@ T1.3%
```

2. general likelihood ranking for single word and idiom tags

   the ranking stems from frequency-based dictionaries, past tagging experience and intuition; for example some dictionaries list the first sense of *odd* as 'strange, unusual' (as in *odd behaviour*), but our experience with spoken data shows that the 'occasional' sense (as in *odd pint*) is more frequent.

3. overlapping idiom resolution

   Normally, semantic idioms take priority over single word tagging, and in some cases our set of idiom templates produce overlapping candidate taggings of the same set of words. A set of heuristics is applied to enable the most likely idiom to be promoted as the top choice. The heuristics take account of length and span of the idioms and how much of a wildcarded template matched in each case.

4. domain of discourse

   This is used to alter the rank ordering on semantic tags in the lexicon and idiom list for a particular domain. Consider the adjective 'battered' which we give three candidate tags (in the following general rank order) 'Calm/Violent/Angry' (e.g. battered wife), 'Judgement of appearance' (e.g. battered box), and 'food' (e.g. battered cod). If the topic of conversation was known to be on fish-and-chip shops, then we can automatically raise the likelihood of the third tag by promoting 'food' (F1) tags.

5. auxiliary/content rules

   A large number of ambiguities are caused by the verbs *be*, *do* and *have.* We have developed part-of-speech templates which make use of the fact that the auxiliary usage collocates with a specific verb form. The same technique is also applied to high frequency words which can be disambiguated easily with short context rules.

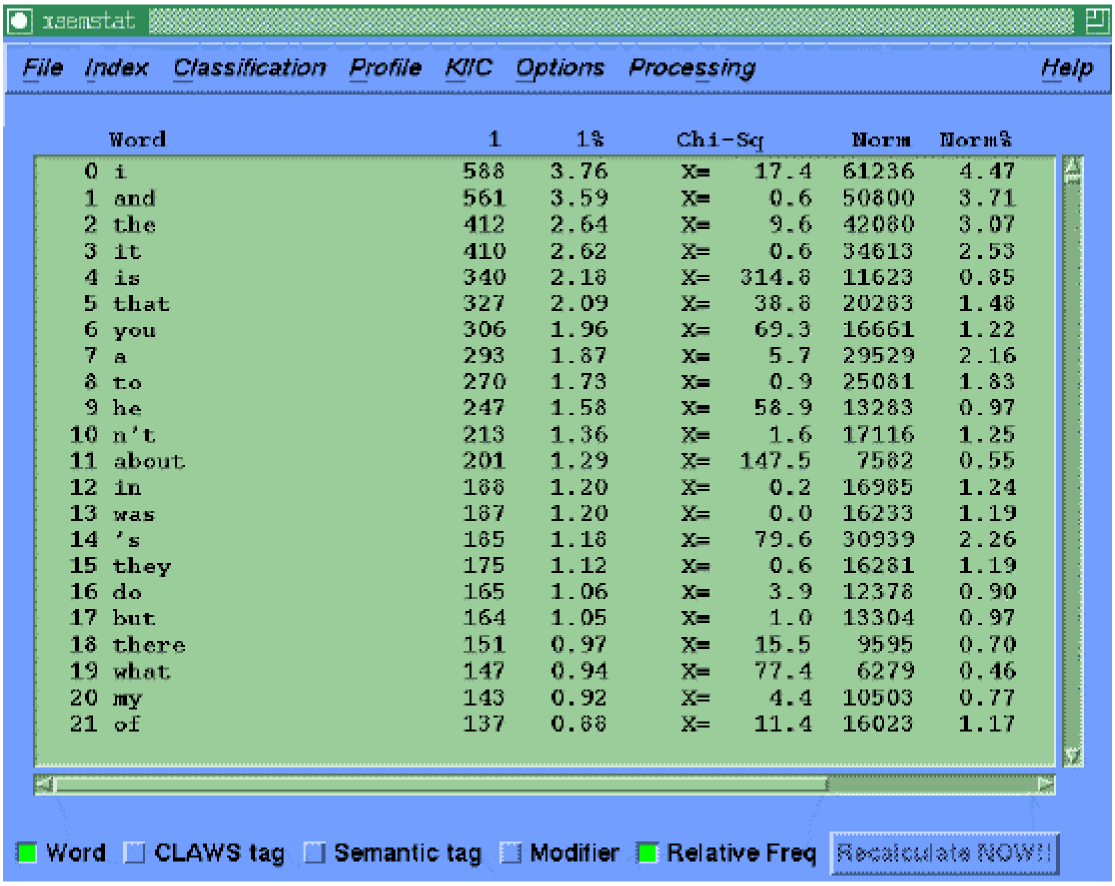
6. proximity disambiguation

   The proximity disambiguation method is still under development. It is based on the assumption that the correct semantic tag for a word in a given context is related to the semantic tags of words in the surrounding context. More particularly, we feel that if a word is generally ambiguous between a set of semantic tags then in a given context we can disambiguate it on the basis of the frequency of co-occurring tags in the surrounding context. For each semantic tag on a semantically ambiguous word in the text, we calculate, using a similarity matrix derived from hand-corrected data, a weight for every other tag which co-occurs with it within a given context window, taking into account the distance of the co-occurring tag from the focus tag and the likelihood of the co-occurring tag in its own context. The highest weight tag then 'wins'.

Retrieval using SEMSTAT

SEMSTAT has been designed within the ACAMRIT project, but can be used (for research at Lancaster) as a stand-alone package with other texts in a variety of different formats. There is a user-friendly graphical user interface (on X-windows) and a character-based terminal version.

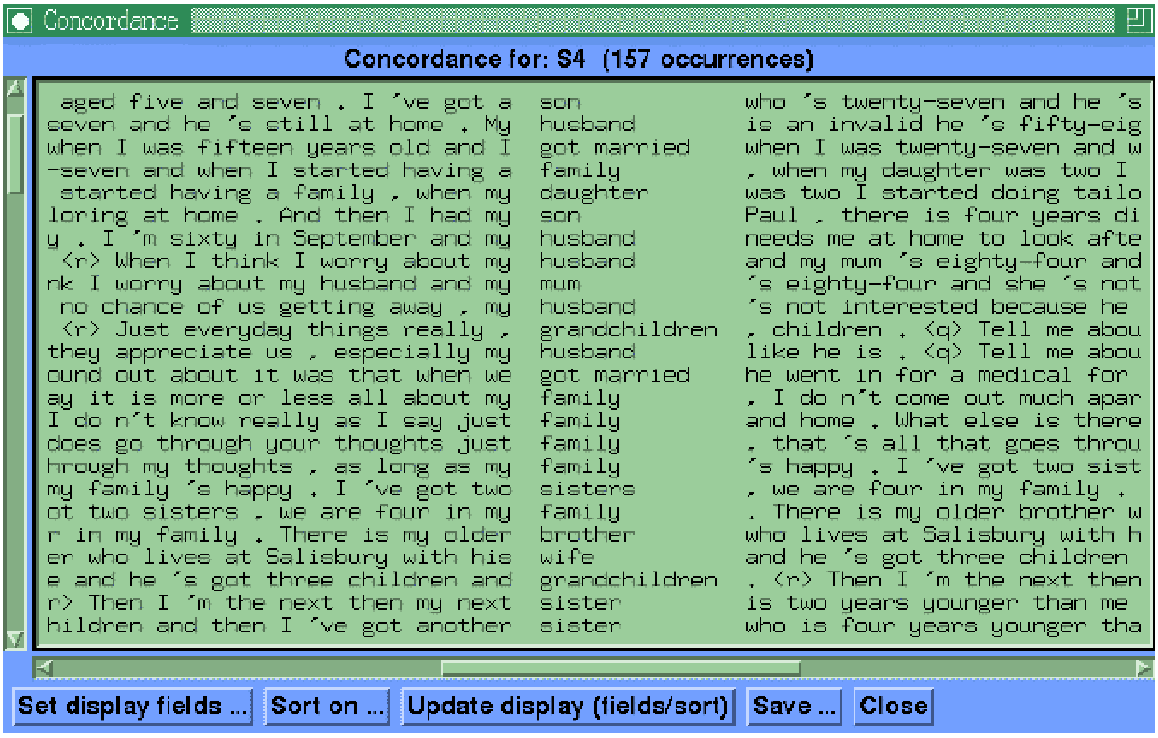
SEMSTAT first displays ACAMRIT data as a semantic tag frequency profile. The user can interactively change the view they see to include other fields such as word, POS tag, linked words (from MATRIX), relative frequency and a dispersion value (which shows how many interviewees mention each word/tag). Each line in a profile displays the χ^2 value that shows which items differ significantly from an expected frequency derived from the normative corpora.



	Word	1	1%	Chi-Sq		Norm	Norm%
0	i	588	3.76	X=	17.4	61236	4.47
1	and	561	3.59	X=	0.6	50800	3.71
2	the	412	2.64	X=	9.6	42080	3.07
3	it	410	2.62	X=	0.6	34613	2.53
4	is	340	2.18	X=	314.8	11623	0.85
5	that	327	2.09	X=	38.8	20283	1.48
6	you	306	1.96	X=	69.3	16661	1.22
7	a	293	1.87	X=	5.7	29529	2.16
8	to	270	1.73	X=	0.9	25081	1.83
9	he	247	1.58	X=	58.9	13283	0.97
10	n't	213	1.36	X=	1.6	17116	1.25
11	about	201	1.29	X=	147.5	7582	0.55
12	in	188	1.20	X=	0.2	16985	1.24
13	was	187	1.20	X=	0.0	16233	1.19
14	's	185	1.18	X=	79.6	30939	2.26
15	they	175	1.12	X=	0.6	16281	1.19
16	do	165	1.06	X=	3.9	12378	0.90
17	but	164	1.05	X=	1.0	13304	0.97
18	there	151	0.97	X=	15.5	9595	0.70
19	what	147	0.94	X=	77.4	6279	0.46
20	my	143	0.92	X=	4.4	10503	0.77
21	of	137	0.88	X=	11.4	16023	1.17

χ^2 values are known to be unreliable for items with expected frequency lower than 5 (see Dunning, 1993), and possibly over estimate with high frequency words. Hence, the results are usually checked using the dispersion value and concordances to take into account the distribution within the corpus.

Within any view the user can double click on the profile to see a concordance of the selected item. It is also possible to display other fields in the concordance window so that the user can see patterns of tagging surrounding a key item.



```
Concordance for: S4  (157 occurrences)

  aged five and seven . I 've got a  son            who 's twenty-seven and he 's
 seven and he 's still at home . My  husband        is an invalid he 's fifty-eig
 when I was fifteen years old and I  got married    when I was twenty-seven and w
 -seven and when I started having a  family         , when my daughter was two I
  started having a family , when my  daughter       was two I started doing tailo
 loring at home . And then I had my  son            Paul . there is four years di
 y . I 'm sixty in September and my  husband        needs me at home to look afte
  <r> When I think I worry about my  husband        and my mum 's eighty-four and
 nk I worry about my husband and my  mum            's eighty-four and she 's not
  no chance of us getting away . my  husband        's not interested because he
  <r> Just everyday things really ,  grandchildren  , children . <q> Tell me abou
 they appreciate us , especially my  husband        like he is . <q> Tell me abou
ound out about it was that when we  got married    he went in for a medical for
 ay it is more or less all about my  family         , I do n't come out much apar
 I do n't know really as I say just  family         and home . What else is there
 does go through your thoughts just  family         . that 's all that goes throu
 hrough my thoughts , as long as my  family         's happy . I 've got two sist
 my family 's happy . I 've got two  sisters        , we are four in my family .
 ot two sisters . we are four in my  family         . There is my older brother w
 r in my family . There is my older  brother        who lives at Salisbury with h
 er who lives at Salisbury with his  wife           and he 's got three children
 e and he 's got three children and  grandchildren  . <r> Then I 'm the next then
 r> Then I 'm the next then my next  sister         is two years younger than me
 hildren and then I 've got another  sister         who is four years younger tha
```

Using a classification scheme based on the information encoded in the file headers a user can select subcorpora and hide parts of the text not of interest (for example the interviewer's questions). The scheme also allows the user to display frequencies for different parts of the corpus alongside each other. The χ^2 value is then used to show items whose frequency distribution across the subcorpora is statistically significant.

Other uses of the system

Apart from market research, the system has other potential applications in linguistics and more generally in the social sciences and humanities: for example, a pilot study of a corpus of doctor-patient interactions has been carried out using ACAMRIT (see Thomas and Wilson, 1996), and its application to the stylistic analysis of written as well as spoken English has been piloted by Wilson and Leech (1993). We are currently using SEMSTAT to compare native and non-native speakers of English in Sylviane Granger's ICLE (International Corpus of Learner English) corpus (see Granger, 1993).

Bibliography

Crowdy, S. (1993), 'Spoken corpus design and transcription', *Literary and Linguistic Computing,* 8(4), 259-265.

Crowdy, S. (1995), 'The BNC Spoken Corpus', in G. Leech, G. Myers and J. Thomas (eds.), *Spoken English on Computer: Transcription, Mark-up and Application*, London: Longman, 224-234.

Dunning, Ted. (1993). 'Accurate methods for the statistics of surprise and coincidence' *Computational Linguistics,* Volume 19, number 1, 61-74.

Garside, R. (1996). The robust tagging of unrestricted text: the BNC experience. In J. Thomas and M. Short (eds.) Using corpora for language research. Longman, London, pp 167-180.

Garside, R., McEnery, A., and Rayson, P. (1993). *Argument Frame Extraction and Term Clustering from an English-French Bilingual Aligned Corpus*. ET10-63 Working Paper.

Granger S. (1993). The International Corpus of Learner English. In J. Aarts, P. de Haan and N. Oostdijk (eds.), *English Language Corpora: Design, Analysis and Exploitation.* Rodopi, Amsterdam & Atlanta, 57-69.

McArthur, T (1981). *Longman Lexicon of Contemporary English.* London, Longman.

Thomas, J., and Wilson, A. (1996). Methodologies for studying a corpus of doctor-patient interaction. In J. Thomas and M. Short (eds.) *Using corpora for language research*. Longman, London, pp 92-109.

Wilson, A. (1991). No, not, and never: negation in a corpus of spoken interview transcripts. *Lancaster papers in Linguistics,* number 73.

Wilson, A. (1993). Towards an Integration of Content Analysis and Discourse Analysis: The Automatic Linkage of Key Relations in Text. *UCREL Technical Paper number 3,* Department of Linguistics, Lancaster University.

Wilson, A. and Leech, G.N. (1993). Automatic Content Analysis and the Stylistic Analysis of Prose Literature. *Revue: Informatique et Statistique dans les Sciences Humaines* 29: 219-234.

Wilson, A. and Rayson, P. (1993). Automatic Content Analysis of Spoken Discourse: a report on work in progress. In: C. Souter and E. Atwell (eds.*), Corpus Based Computational Linguistics*. Amsterdam: Rodopi. pp215-226