The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0001-253X.htm

Semantic targeting: past, present, and future

David Crystal
Department of Linguistics, University of Bangor, Bangor, UK

Received 20 January 2010
Revised 14 May 2010
Accepted 28 May 2010

Abstract

Purpose – This paper seeks to explicate the notion of "semantics", especially as it is being used in the context of the internet in general and advertising in particular.

Design/methodology/approach – The conception of semantics as it evolved within linguistics is placed in its historical context. In the field of online advertising, it shows the limitations of keyword-based approaches and those where a limited amount of context is taken into account (contextual advertising). A more sophisticated notion of semantic targeting is explained, in which the whole page is taken into account in arriving at a semantic categorization. This is achieved through a combination of lexicological analysis and a purpose-built semantic taxonomy.

Findings – The combination of a lexical analysis (derived from a dictionary) and a taxonomy (derived from a general encyclopedia, and subsequently refined) resulted in the construction of a "sense engine", which was then applied to online advertising, Examples of the application illustrate how relevance and sensitivity (brand protection) of ad placement can be improved. Several areas of potential further application are outlined.

Originality/value – This is the first systematic application of linguistics to provide a solution to the problem of inappropriate ad placement online.

Keywords Semantics, Advertising, Electronic media

Paper type Conceptual paper

1. Historical background

Semantics began as a branch of linguistics, the science of language. Indeed, the word "science" is used in its original definition: the French philologist Michel Bréal, who introduced the term in the 1890s, defines it as "la science des significations" – the science of meaning in language. It came to be seen as a "level" of linguistic investigation, alongside phonetics, phonology, morphology, and syntax, in such seminal works as Leonard Bloomfield's (1933) *Language*, but the abstract and indeterminate nature of "meaning" meant that it remained a neglected branch of linguistics for many decades. The first full-scale linguistic treatment was John Lyons' (1977) two-volume *Semantics*, now regarded as a classic statement of the "state of the art" within linguistics and linguistic philosophy. In the meantime, in the absence of a linguistic characterization, other fields found the notion of semantics useful and began to employ it in individual ways.

The philosopher Charles Morris (1946) gave semantics a more general interpretation, defining it as the interpretation of signs in general – signs here being used in an abstract sense to include everything that conveys information. It therefore included facial expressions, bodily gestures, road signs, railway signals, and other non-linguistic systems. Also in the 1940s the term achieved certain notoriety in popular usage, where "it's just semantics" began to refer to an irritating or pointless quibble. Psychologist Charles Osgood (1953) took the term in a different direction, referring to

Aslib Proceedings: New Information Perspectives
Vol. 62 No. 4/5, 2010
pp. 355-365
© Emerald Group Publishing Limited
0001-253X
DOI 10.1108/00012531011074627

the judgements people make about words, and devising a system of rating scales which he called a "semantic differential" – whether words are judged as strong/weak, good/bad, active/passive, and so on.

Sometimes the term is narrowed, as when it appears in medicine with reference to a clinical syndrome – "semantic aphasia" – where the patient loses the ability to use words after brain damage. Sometimes it is broadened, as when Alfred Korzybski developed "general semantics" in the 1930s as a method of enabling people to avoid the ideological traps built into language. And, of course, it has achieved one of its widest extensions in the notion of the "semantic web", where it includes all concepts and relationships within human knowledge. There could be no broader definition of "semantics" than the one we encounter in the "semantic web", and no definition that is further away from the original linguistic intention. The semantic web, in the words of its originator, will evolve "without relying on English or any natural language for understanding" (Berners-Lee, 1999, p. 203).

I give this brief historical background to make it clear that anyone who claims a "semantic" solution to a problem in knowledge management has to be viewed with a certain amount of caution. What sort of semantics is being propounded? The term has achieved a buzz word status these days, with many companies and approaches calling themselves "semantic". It must not be assumed that they are all talking about the same thing, or focusing on the same aspects of language. Ogden and Richards (1923) wrote a book called *The Meaning of Meaning*, which explored sixteen senses of that term. Today a book called *The Semantics of Semantics* would have to deal with far more.

2. Linguistic semantics

The one thing the linguistic approach to semantics has taught us is that the meaning system of a language is immensely complex. Meanings are notoriously difficult things to pin down, as illustrated by such well-investigated notions as connotation, collocation and fuzziness. Even an apparently simple notion like "oppositeness" turns out to be quite complicated. To take just one topic from the structural semantics literature as an illustration: quickly is the "opposite" of slowly, and single is the "opposite" of married, but we can say very slowly and more quickly whereas we do not say very single and more married (except when joking). There are evidently two kinds of opposite here – what are often called "gradable" and "non-gradable" antonyms, respectively. And other sorts of oppositeness have been identified. Linguistic semanticists have been working on semantic questions of this kind for over half a century, both formally ("formal semantics", a field which overlaps with logic and mathematics) and descriptively (leading to applications in dictionary-compilation and translation); their unanimous conclusion is that this is probably the most complex area of linguistic enquiry, and one which is still in its early stages of development. It is a conclusion that proponents of the semantic web need to take into account.

It is not that all areas of meaning are equally complex. There are some clear areas: cases which are fairly straightforward, in that the meanings are relatively concrete, discrete, and determinate and enable clear statements to be made quite quickly (such as the terms which identify the elements of a kinship system, for example "male" versus "female", "parent" versus "child"). But these are far outnumbered by those where the meanings are relatively abstract, fuzzy, and subject to variation between people and cultures. Indeed, even kinship is not immune to cultural variation: what counts as a

"brother" or "uncle" in one culture may be hugely different from what is a "brother" or "uncle" in another. Estimates vary greatly, but the amount of cultural distinctiveness within a language as reflected in its lexicon (i.e. those elements in the vocabulary of one language which carry cultural connotations that are not easily translatable into another) is anywhere between 5 and 25 per cent (depending on how closely the languages are related to each other).

There are many other possible illustrations of the kind of complexity we need to anticipate in projects such as the semantic web. For example, reference to any dictionary will show the way languages are essentially polysemic, i.e. the words have more than one meaning. The existence of huge amounts of monosemic scientific or technical vocabulary must not be allowed to obscure this point. Over 70 per cent of the million plus words in English are technical terms, many of which (though by no means all, as any science dictionary shows) have an agreed single meaning within the relevant science, but these are not the items which are of primary concern in investigating the meaning-system of a language. If we take a typical "college" or "concise" dictionary of the words in routine daily use – such books contain around 100,000 headwords or so – we would find that the average number of senses per headword is 2.4 (for English), and rapidly rising (because of the way everyday words are being assigned to internet domains with increasing frequency); in an unabridged dictionary, such as the large *Oxford English Dictionary*, that figure would be much higher. How to handle a rapidly evolving polysemy has to be at the heart of any realistic project which aims to deal with the way meaning is present on the web.

Until recently, very little recognition had been given to the importance of linguistic semantics in solving problems of search, navigation, and classification in computer-mediated communication. And even today, problems abound. Searches on Google, for example, still produce unacceptable levels of irrelevance and incoherence. It remains an everyday experience to type a query-term (such as bridge) into a search-box and receive an enormous variety of miscellaneous hits, thereby illustrating the distance that still has to be travelled before search achieves relevance levels which are intuitively satisfactory. In the world of advertising, the disasters happen daily. A CNN page on a street stabbing in Chicago is accompanied by ads saying "buy your knives here". A German page about a tour of Auschwitz is accompanied by ads from one of the German power companies advertising cheap gas. It takes only a five-minute trawl on the web to bring to light several such examples.

3. Semantic targeting

The aim of my work in applied semantics over the past 12 years has been to find a solution to the problems of irrelevance and insensitivity caused by clashes of this kind. The diagnosis, of course, is easy: it is clear what has happened, in cases such as the CNN report. The primitive algorithm employed by the ad-placement company has found the keyword "knife" a few times on the web page and linked this with the ads in its inventory which also used the word – namely, cutlery ads. But the effect was not as intended. Plainly, a keyword approach will never work, because it fails to take into account the ambiguity presented by most words in a language.

An apparent solution, which I worked with for a few years, was to put words (technically, lexemes) into context. The argument went like this: the word "knife" in the CNN report will be accompanied by such other words as "police", blood, "body" and

"murder". The word "knife" in the ad inventory will be accompanied by words such as "fork", spoon, "cup" and "plate". By taking those other words into account, "knife" will be disambiguated into its "weapons" and "cutlery" senses; indeed, it is possible to use this "contextual semantics" to distinguish senses in this way. The approach had its adherents, and the approach called "contextual advertising" was one of the outcomes. But a contextual approach is not the solution, for it captures only part of the content of a page. The CNN page was about a street stabbing, certainly, but it was also about a range of other issues, such as street safety in Chicago, policing methods, and citizen protection. This is typical. Most web pages (forums, blogs, etc.) are multi-thematic. There is a natural intuitive tendency to think that what a page is "about" is identified by its headline and first paragraph. But read down to the bottom of a page, and other themes soon come to the fore. It is very rare indeed to find a web page which has just a single theme. So if we want to see ads down the side of a web page which are relevant to the content of that page, it is essential to have a means of analysing the content of the entire page. It is this whole-page content analysis which over the past five years I have come to call "semantic targeting".

Semantic targeting has several possible areas of application in addition to online advertising: search engine assistance, to improve the relevance of search results; e-commerce, to improve the accuracy of online enquiries; automatic document classification, to facilitate the retrieval of information in large electronic databases; and internet security, to monitor sensitive or dangerous online content. In each of these areas, two types of information are required for a semantic targeting procedure to work. Pages have to be lexically analysed and they have to be thematically categorized. Where do the words and categories come from?

4. Establishing the lexicon

If you are interested in, say, the weather, and you type into a search engine the word "depression", you will get millions of hits. But the first page of results will give you little or nothing about the weather: what you will get is a page of results offering you advice and drugs about how to cure your state of mind. How can this situation be improved? It might be thought that simply increasing the number of search terms will solve the problem: not always. Because of the way search engines typically work, increasing the number of search terms can bring an increased diversity of results, especially as the enquiry becomes more abstract in character. Some relevant results will be included, of course, but they can be hidden within a welter of irrelevant hits. And in any case, thinking up exactly which search terms produce the best results is not always an easy matter.

Plainly the problem illustrated by "depression" arises from the polysemic character of that word. In fact it has four main meanings, each of which relates to a knowledge category: mental state (psychiatry), bad weather (meteorology), poor economy (economics), and a dip in the ground (geology). If it were possible to devise a semantic filter which distinguished these meanings, the problem would be solved. In the prototype scenario I developed in the 1990s, a search-engine user would type in "depression" and up would come a menu which would say "Which sense of depression do you require?" and the four contexts would appear. The user would click the one desired, a semantic filter would operate, and only the hits related to the chosen

category would be allowed to appear. The question accordingly was: how to provide the semantic content for the filter?

The answer was simple, but time-consuming. To continue with the "depression" example: if enquirers wanted to see only web pages to do with the weather, then all they had to do was predict which weather words were likely to appear on the page; if "depression" appears on a page which also contains such items as "rain", low "pressure", and "windy", then the page is likely to be about meteorology rather than psychiatry or economics. Conversely, if "depression" appears on a page which contains such items as "symptoms", illness, and "prozac", then it probably is not going to be about the weather. So the question became: how many items are there in the English language which are available to users to talk about the weather – or economics – or psychiatry – or geology? If we can predict what all of these are, then the content of the filter (for these categories) would be comprehensively defined.

The same solution is needed in the case of the advertising example. If we want to place ads appropriately in relation to the diverse thematic content of a page, then we must predict the words that identify each theme. It is the same question: how many words are there in the English language which are available to talk about homicide, or Chicago, or policing? Or, – to take other themes of interest to advertisers – refrigerators, cars, mobile phones, and holidays in Bermuda? If we can predict these words, then the content of web pages can be comprehensively classified.

We can generalize. We need to be able to predict the category-specific lexical content for any desired knowledge category. It is not enough for a system to look at the pages which have already been written about, say, refrigerators. We need a system which can say in advance what words are likely to turn up on any page about refrigerators which has yet to be written. How many refrigerator-specific words are there in English (or other languages)? If we try to answer this question off the cuff it turns out to be very difficult. A few words come to mind (fridge, freezer, shelf, door, cold...) and then our mind goes blank. We know there must be many more, but what exactly are they? We need comprehensiveness. Where is it to be found?

There is one place where all the lexical items in a language are gathered together: a dictionary. So the research task was clear: I had to work my way through all the content-specific items in a dictionary (i.e. excluding grammatical words such as "the" and "of", and semantically "light" items such as "make" and "get") and assign each sense of each item to a semantic category. I used *Chambers 21st Century Dictionary* as my basic text, and supplemented this by specialist works as required and by internet searches (which provided most of the proper names – brand names, place names, and the like) that would not normally be included in a dictionary. At the end of this exercise, I did indeed know just how many words there are in English for refrigerators, or cars... or anything.

It took three years and a team of 40 part-time lexicographers to complete the construction of the lexical database – or, I should say, to complete a first pass, for this kind of task is never-ending. New terms are constantly being introduced into a language, and they have to be added. For example, in 2000 the set of lexical items relating to Iraq did not include the phrase "weapons of mass destruction". This had to be added in 2003. Or, to take a more commercial example, as new models of motor-car come on the market, their names or model designations have to be incorporated. All

semantic databases have to be maintained in this way, for language and society never stop changing.

The task took so long because it involved several linguistic considerations:

- *Semantically*: the lexicographers had to identify individual lexical items, their senses, and any high-frequency collocations.
- *Grammatically*: they had to identify compounding alternatives (is it "flowerpot" or "flower pot"?) and all inflectional variations (such as singular and plural of nouns).
- *Sociolinguistically and stylistically*: they had to identify formal and informal variants (e.g. "television" and "telly"), regional differences (chiefly American versus English variants, such as "boot" and "trunk", color and "colour"), and within-region spelling alternatives (e.g. "judgment" and "judgement").
- *Each item had to be weighted*: to indicate its value as an identifier of a category. For example, the word "quarterback" is a high-value identifier because it occurs only in the category of American Football. "Depression" is a medium-level identifier because it turns up in at least four categories, as we have seen. And "country" is a low-level identifier, because it turns up in dozens of categories (such as within travel, politics, and the environment). All items were weighted on a scale from 1 to 100.

Despite all these variables, the number of items in a category is not as large as one might think – for higher-order categories (such as "motor vehicles") there might be as many as five hundred; for lower-order ones (such as a specific brand of car) there might be less than a dozen. In its current version there are around 3,000 key-worded categories in the taxonomy, with an average number of items per category of 104.7, and a range from 5 to 514. The underlying lexicon is a little in excess of 300,000 items.

The entire approach, along with the software which drives it, I came to call a "sense engine". Note that a sense engine makes no assumptions in advance as to what a page is going to be about. The page is tested against all 3,000 knowledge categories, to see which ones are relevant. And there are often surprises. Thus, a report on a win by a tennis star at Wimbledon was rightly classified as tennis; but the sense engine also said the page was about cars and dating. Only by reading the whole page did this become clear. After reporting the tennis win, the writer of the article went on to talk about the star's taste in cars and women. From an advertising perspective, this example illustrates two issues, which can be summed up in the words "relevance" and "protection". Semantic targeting offers an opportunity for automobile firms, for example, to place a relevant ad on the page; it also offers an opportunity for brand protection, as it alerts agencies to a possible sensitivity, for they may have clients that do not want their ads to appear on a page which contains any kind of sexual reference.

5. Establishing categories

It is clear that there are two dimensions to semantic targeting. The sense engine needs a set of knowledge categories and it needs a set of lexical items to characterize each category. The dictionary project provided the lexical items. Where did the knowledge categories come from?

In 1986 I was invited to be editor of an encyclopedia project which had been initiated by a joint venture of Cambridge University Press and W & R Chambers. This was an exercise in general reference publishing. My role was to plan the structure of the work, find contributors, edit their contributions, and bring the project to fruition, which I eventually did in 1990 when the first edition of *The Cambridge Encyclopedia* appeared (Crystal, 1990). To manage the large quantities of data involved, I had to devise a classification system. Each encyclopedia entry was classified into a number of knowledge categories. The entry on "Winston Churchill", for example, was classified with reference to politics, journalism, art, literature, and so on, reflecting the many aspects of his life. It thus proved possible to find in our database "all the novelists", or "all the 19th century novelists", or "all the 19th century French novelists", and so on. At the end of the encyclopedia project, my taxonomy contained some 500 categories.

In 1995 everything changed. Owing to a new policy within CUP, a decision was made to divest themselves of my operation, and they sold the entire encyclopedia portfolio to a Dutch IT firm called AND. AND were not so much interested in the encyclopedias as in the associated classification system. They could see the potential of our taxonomy for improving results coming from the search engines which were around at the time, such as Excite, Lycos, and AltaVista. My main role between 1996 and 2001 was to develop the taxonomy to make it work on the internet. While the categories I had devised for CUP (within literature, sociology, earth sciences, philosophy, and so on) were still relevant, they did not tell the whole story. Indeed, much of the content found on the internet was missing. You may find this hard to believe, but sex, fast cars, and travelling to Las Vegas did not figure largely in the editorial remit I had received from the oldest press in the world. The taxonomy grew enormously, as a result, from some 500 categories to around 1,500.

In 2001 everything changed. Owing to an over-ambitious acquisitions policy, at a time when dot.coms were failing everywhere, AND went into liquidation, and my editorial office was closed down. Determined not to waste what was by then 15 years of work compiling the encyclopedia database, as well as a taxonomy which contained considerable potential for application (and which had been granted UK and US patents), an ex-AND colleague and I decided to set up our own company to develop possible products. We called it Crystal Reference Systems, and from 2001 to 2006 we continued to publish encyclopedias – this time for Penguin Books – and developed the software technology which was required in order to put the taxonomy to work. But in which direction should we go? Document classification? Search engine assistance? Internet security? E-commerce? Online advertising?

In 2006 this was decided for us. The contracts we needed did not come quickly enough to enable us to survive on our own. We needed a partner, and in that year we were bought up by a European-wide online advertising business called Adpepper Media. The consequences of this move were that our focus then became exclusively directed towards developing the sense engine in relation to online advertising, and a whole new set of goals arose. The taxonomy had to grow in fresh directions to meet advertising priorities: categories had to be devised for hundreds of commercial topics, such as refrigerators, BMWs, and credit cards. The number of categories in the taxonomy grew to 3,000, as already mentioned, and the implementation of the sense engine led to a commercial product called iSense, which ensures that an ad is appropriately placed on a particular page.

The focus on advertising brought a second goal to the fore, in addition to relevance, which can be summed up by the word "sensitivity". In the ad world, it is also crucial to ensure that an ad is not placed on an inappropriate page – an ad for children's clothing on an adult porn site, for example. Companies get very angry when this inadvertently happens. There are several internet domains which raise problems for advertisers. For example: sites to do with smoking, drinking, gambling, weapons, pornography, and nudity; sites which present extreme views to do with politics or religion; sites which introduce a great deal of swearing. Most advertisers (other than those which specialise in such areas, of course) do not want their ads to appear on such sites. How can misplacement be avoided? The arrival of Adpepper made this a new and immediate priority.

The sense engine, which had previously been used in a positive way, to include as much as possible, now had to be adapted to exclude. The procedure was the same as before. Each of the dangerous categories had to be explored to identify the set of lexical items which characterized them. Conventional dictionaries were of limited value in this respect: the full range of colloquial pornographic vocabulary, for example, does not appear in most dictionaries. It was an interesting few weeks for me, I must admit, as I searched porn sites not looking – as I repeatedly had to assure my wife – at the bodies of the hunks that were there but at the words used to describe their bodies. But we linguists are made of strong stuff, and I soldiered on. The result was a filter (which I called Sitescreen) which can identify sensitive sites so that advertisers can avoid them.

The arrival of Adpepper brought a third goal to the fore, which again changed the priorities of our research: "localization". It is all very well having a database in English, but what about other languages? Adpepper had branches in 12 European countries, and the need to provide ad relevance there was as strong as in English-speaking countries. The need to translate the database into their languages suddenly became urgent. It was going to be a huge task. The 300,000 lexical items had to be not just translated, but localized. It is possible to make a straightforward translation from English into these languages for something like three-quarters of the vocabulary. The meteorological sense of "depression" in English will neatly equate to a corresponding word in French, German, and so on. But in around a quarter of cases, there is no direct one-to-one translation, partly for linguistic reasons and partly for cultural reasons. Semantic mismatch is a familiar issue in translation theory, summed up by the popular saying, "The French (or whoever) have a word for it". Cultural mismatch can be illustrated by the task of translating the names of popular cigarette brands or drinks, which vary from country to country, or by the task of finding what the cultural equivalents are for political or minority groups, especially when used in insulting ways: what is the French equivalent of "Paki", for instance? This is time-consuming and difficult work, and it has taken about three years to produce iSense and Sitescreen databases for the languages initially selected for translation (German, Danish, Dutch and French).

One complication is that the goals are continually changing. The ultimate advertising goal is to place ads on web pages so that they relate as closely as possible to the content of the page. If the page is about Britney Spears, then once on a time it was enough simply to ensure that the ads were about music, rather than about, say, weapons (spears). Then the demand narrowed: the ads had to be about popular music, and not classical music. Then the demand narrowed further: the ads had to be about

Britney Spears as such. The most recent demand requires yet a further narrowing: some advertisers only want their Britney Spears ads to be placed on pages which say nice things about her. If a new album is given a bad review, they do not want to be associated with it. The same point applies to commercial goods. A firm like Hotpoint does not want to advertise on a web page or forum which says that Hotpoint washing machines are rubbish. So now there is a new goal, which can be summed up in another single word: "sentiment". Can one identify the sentiment of a web page? It is indeed possible, but it requires another lexicographic trawl – this time identifying all the words in a language which express positive and negative attitudes. (Out of interest: there are c.1,500 words in English for positive attitudes and c.3,000 for negative ones.) This linguistic task is trickier. Compare: "Britney Spears' latest album is rubbish" versus "Britney Spears' latest album is by no means rubbish". The reversative force of negative words has to be taken into account. And there are several other syntactic considerations involving word order and the use of intensifiers (such as *very*). An originally lexical exercise now takes on a grammatical dimension, and the research is forced to move in the direction of the kind of issue that has long been a central concern of natural language processing.

6. Future applications

The industrial world is always changing the goalposts, as it responds to what is perceived to be the needs of the customers. Within the course of ten years, my industry-inspired research priorities have changed four times, as summarised in my watchwords: relevance, sensitivity, localization and sentiment. Nor is this the end of the story. A recent focus of the advertising industry is behavioural profiling. Here the question is no longer "Do people like Britney Spears?" but "Do *you*, John Doe, Mary Smith, like Britney Spears?" Is it possible to tell, from an analysis of your blog, or your page on Facebook or wherever, what your interests are to the extent that a highly personalized advertising campaign can be targeted at you? This is not an issue for semantics. The social or ethical issues involved go well beyond the linguistic. It is a complex arena, which has raised controversial questions over privacy. Whether behaviours are profiled or not, pages still need to be semantically categorized. Semantic targeting complements behavioural approaches.

Another future area of application for semantic targeting is in relation to internet security. Is it possible to identify a dangerous conversation on the internet – for example, e-mails between terrorists or fraudsters, or paedophile grooming in chatrooms? Several internet and mobile phone organizations have expressed their concern. Yes, it is possible, if the data are available to enable the research to be done. The same sense-engine methodology can be used, but adapted now to handle a dynamic situation – an ongoing conversation. It is not possible to tell, from a single sentence in a chatroom, whether the speaker is an adult masquerading as a child. But it is certainly possible to tell that a dangerous conversation is taking place if one plots a series of "leading" sentences over time. Once one has identified the salient lexical items (e.g. in a paedophile context, such words as "clothes" and "wearing"), it is possible to develop a lexical filter which can provide a cumulative score. An innocent conversation will score low; a dangerous one will quickly score high.

I devised a procedure of this kind, called Chatsafe, as part of our activities in the early 2000s, but it proved impossible to take it forward. Why? Because to test it,

I needed access to real paedophile conversations, and apart from a sample or two provided by child protection agencies on the web, these proved impossible to get hold of. I contacted the police, the Home Office, and others who had tried to research this area, and the message was the same. If I continued to explore this domain without top-level clearance, I risked arrest! But I could never discover how I could get such clearance. And I heard horror-stories – such as the external examiner who was interviewed by police simply for reading a PhD thesis on paedophile activity. It also turned out that for a university even to send such a thesis through the post to an examiner was risking a criminal prosecution. Can such research be done at all, I wonder? The need is there and a possible linguistic solution is available. But in the absence of testing, it remains on the shelf.

Has semantic targeting any limitations? It requires pages with textual content, so if a page is predominantly image-based, the amount of text available for lexical analysis may be quite small – though usually a scrape of the metatags in a page provides enough data for a classification to be made. The same point applies to video files. If a video file can be found through a web search – a difficult enough task in itself – then the underlying tags which enabled this to happen may be enough to allow it to be classified accurately. If there is speech content, the problem is likely to be easier, as, thanks to recent advances in automatic speech recognition, a speech-to-text algorithm could provide a transcription of sufficient accuracy to enable an accurate classification to be made. The beauty of semantic targeting is that, because it deals with all the content words on a page, it can achieve accurate results even if some of those words are unanalysable. (Similarly, the procedure works well even if there are occasional misspellings; though a fuzzy word recognition algorithm, such as is used by Google to identify mistypings, would certainly be a useful addition in due course.)

Semantic targeting is as good as the underlying linguistics, so that if there are types of expression which have so far not received a sophisticated semantic analysis, these will not be handled well. Texts involving a great deal of sarcasm, irony, or figurative expression are illustrative. If a text were significantly figurative, as in some kinds of poetry, then it would not classify well at present. Fortunately, in the advertising world, figurative expression is limited in scope. It may well be that, in the middle of a football page, the writer describes rain causing the players to "slide about like ducks on ice", but this alone would not be enough to cause the sense engine to misclassify the page as ornithology. The majority of the page would be about football, with the words used literally, and an appropriate classification would be made; having said that, this is an area which has been very little explored.

Also little explored is the application of semantic targeting to dynamic text. For the sense engine to work, it needs to capture a body of text. It is not clear how this is to be done if the text is mobile, as in some Flash animations. Nor is it clear how best to handle a page where people are continually adding fresh material, as in a forum or a social networking site. The only way I've done this hitherto is to sample a page or site at arbitrary time intervals, and ignore its intervening textual evolution, but plainly a more sophisticated approach is desirable.

Is semantic targeting the final frontier? No. I believe there is another perspective which will one day become critical: pragmatic targeting. Pragmatics is a recently developed branch of linguistics which studies the choices people make when using language: why does one say or write X rather than Y? What motivates people to use

language in a particular way? What are the intentions behind a particular use of language? In the present context, pragmatics is needed to answer such questions as: Why is a web page written in the way it is? What is its purpose? One web page may be written purely to provide information. Another may be to persuade people to buy something. Another may be to influence opinion. Another may be to inflame opinion (as in some racial sites). There are many possible intentions and outcomes, which will be reflected in the language and layout of the page. To take a simple example, if the page is about buying, it will contain an area with "go to basket", or suchlike. A pragmatic perspective will one day be as important as semantic targeting is now, but it is a long way off, for linguistic pragmatics is still in its infancy. In the meantime, semantic targeting has shown it can provide analyses which are of immediate application in the real world. It is, it seems, protecting brands, avoiding embarrassments and making money. For an applied linguist, that is as good as it gets.

References

Berners-Lee, T. (1999), *Weaving the Web*, Orion, London.

Bloomfield, L. (1933), *Language*, Holt, Rinehart & Winston, New York, NY.

Crystal, D. (Ed.) (1990), *The Cambridge Encyclopedia*, Cambridge University Press, Cambridge.

Lyons, J. (1977), *Semantics*, Cambridge University Press, Cambridge.

Morris, C.W. (1946), *Signs, Language and Behavior*, Prentice-Hall, Englewood Cliffs, NJ.

Ogden, C.K. and Richards, I.A. (1923), *The Meaning of Meaning*, Routledge & Kegan Paul, London.

Osgood, C.E. (1953), *Method and Theory in Experimental Psychology*, Oxford University Press, London.

Corresponding author
David Crystal can be contacted at: davidcrystal1@googlemail.com

To purchase reprints of this article please e-mail: **reprints@emeraldinsight.com**
Or visit our web site for further details: **www.emeraldinsight.com/reprints**