Searchlinguistics

Searchlinguistics is the application of linguistics to analyse and alleviate the problems that arise when people try to provide or obtain information through online search. The term is not yet standard, but the problems are well-recognized, and the demand for solutions is pressing. Several types of e-activity are involved, including the following.

(1) Online advertising, where the requirement is to ensure that ads which appear on a website are relevant, focused, and sensitive to the content of a page.

(2) E-commerce, where the requirement is to obtain data about specific products in online catalogues.

(3) Search engine assistance, where the requirement is to receive accurate, relevant, and up-to-date hits.

(4) Automatic document classification, where the requirement is to find all files which deal with a specific topic or combination of topics.

(5) Internet security, where the requirement is to identify undesirable activity on a site, such as paedophile activity in child chatrooms.

There has been variable progress in each of these areas, but all present difficulties arising out of the lack of a linguistically sophisticated frame of reference, as the following attested examples illustrate.

The problems illustrated

In relation to (1): A CNN page reported a street stabbing in Chicago, and the ads down the side of the screen said 'Buy your knives here', 'Get the best knives on eBay'. A page on a German website described heritage tours of Auschwitz, and an accompanying banner ad from a German energy company advertised cheap gas. The advertising industry is keen to develop more sophisticated methods of targeting ads on pages in order to gurantee relevance and avoid insensitive misplacement. Companies are naturally upset when their ad agency places their products on irrelevant pages or in embarrassing locations. It harms their image and loses sales.

In relation to (2): An enquiry for 'shampoo' (where the enquirer meant 'hair shampoo'), made via a search box on an online retail site, received a cluster of responses mainly about carpet shampoo and car shampoo. Another, to an online bookstore, asked for books by 'David Crystal'; among the list of books received were some by the linguist and some by a Scottish poet with the same name. There was no way of determining which book belonged to which author. The confusion was compounded by the 'further information' provided: 'people who bought this book', said the text accompanying a book written by the Scottish poet, 'also bought the following...' - listing several books by linguists. With a really common name, such as various writers called John Smith, the possibilities for confusion are legion.

The retail industry is keen to develop more customer-friendly methods of interrogating online databases, but finds it difficult to anticipate all the factors which impede good communication. An example is the inadequate specification of variant forms. A search for *mobile phones* on an electronic equipment retail site received the implausible response: 'we have no mobile phones'. Repeated attempts using various lexical, grammatical, and orthographic variants (e.g. *mobile-phone*, *mobile phones*, *cellphones*, *cell phones*) all received the same negative response. Eventually it transpired that the only search term the software recognized was *cellular phones*. Faced with such e-incooperativeness, many people would not have the patience to continue their enquiry, and a sale would be lost.

In relation to (3): A request from a search-engine for information about *apples* (where the enquirer had in mind the fruit) produced several million hits, but all the results on the opening page were about computers and the Beatles, including several results which were seriously out-of-date. An economist who typed in *depression*, expecting information about the financial climate, was swamped with results to do with mental health. Improving the relevance, accuracy, and up-to-dateness of search queries, without making the user do all the work (e.g. by adding extra search words or scrolling through pages of hits), is a continual goal of search engine companies, especially those trying to compete with Google. From the enquirer's point-of-view, the aim is to save time and obtain the most meaningful hits. From a website-owner's point-of-view, the aim is to achieve a high ranking for the site in any set of results.

In relation to (4): A lawyer failed to find all the documents in a database relating to a case in Bosnia and Herzegovina because he searched only for the name of the country in that orthographic form and failed to search for *Bosnia-Herzegovina*, *Bosnia & Herzegovina* and other alternatives. Another found himself flooded with unwanted documents because a search for *New Mexico* also brought in material from Mexico, New York, and other locations containing the word *New*. In the first case, important information relating to the precedents in a case was not retrieved. In the second, a great deal of time was wasted, as the lawyers had to read through a great deal of material before discovering that a document was irrelevant.

In relation to (5): A newspaper article reported the story of a teenager ending up in a dangerous situation having agreed to meet offline an apparently innocent contact made during a chatroom conversation. The contact turned out to be a male predator. Several companies are now concerned to find ways of identifying potentially dangerous content within the discourse of chatrooms and social networking sites. The dangers have increased following the increased provision of content via mobile phones. Parents at least had the opportunity to monitor online activity when this took place through the home computer, but this opportunity is lost when the contact is made directly to a child's mobile. Similar issues arise in relation to the use of the internet to plan terrorism, fraud, or other criminal activities.

The problems analysed

To a linguist, all these problems have one feature in common: they do not take into account the ambiguity inherent in the use of language.

In relation to (1), the software has found the lexeme *knife* appearing several times in the news report, assumed that this was what the page is about, and looked for the same word in the available ad inventory. It ignores the fact that *knife*='weapon' and *knife*='cutlery' are very different linguistic entities. If the software had looked at the linguistic context surrounding the word, it would not have made the erroneous identification: *knife*='weapon' would be accompanied by such lexemes as *murder*, *blood*, and *police*; *knife*='cutlery' would be accompanied by such lexemes as *fork*, *spoon*, and *plate*. A similar ambiguity lay behind the gas example: *gas*='method of killing' vs *gas*='source of domestic energy'.

In relation to (2), the various contexts in which the word *shampoo* appears were not distinguished because the enquirer had not anticipated the ambiguity by using a more specific search term. In this example, the alternatives are easy to see, as *shampoo* is a concrete term with few senses. In the case of words which have many senses, or which are more abstract in meaning, such as *depression*, it is not always obvious how to express a search in such a way that all the unwanted contexts are excluded. In the case of the Amazon authors, the problem arose because of the lack of an appropriate authorial classification in terms of either biography or subject-matter. In the *mobile phones* example, the factors being ignored were to do with British and American English, grammatical number, and orthographic conventions.

In relation to (3), once again the various senses of a word (*apple*, *depression*) were not being distinguished. Plainly the problems arose from the polysemic character of the words - in the case of *depression*, failing to separate the senses relating to the knowledge categories of mental health, meteorology, economics, and geology. It might be thought that simply increasing the number of search terms will improve the relevance of hits in relation to enquiries: this turns out to be not always the case. Because of the way search engines typically work, increasing the number of search terms can bring an increased diversity of results. More relevant results will be found, but they can be hidden within further irrelevant hits. And in any case, thinking up exactly which search terms produce the best results is not always easy. Adding *deep* to *depression*, for instance, will not resolve the ambiguity.

In relation to (4), we see the ambiguity problem in relation to place names. A glance at any gazetteer will show that the same name (*Lancaster*, *Newtown*) can turn up dozens of times in various countries. The orthographic variation encountered in the Bosnia example is just one variable, made more difficult when accents (often ignored by the software) are part of the words. Not all searches can be made successful by the simple expedient of adding an extra locator (e.g. *Lancaster + UK*). For example, it is not a straightforward matter to frame a search so that it finds only entries on the state of *New Mexico* while ignoring all entries relating to the country of *Mexico*.

Ambiguity also arises in relation to (5), but in a different way - involving the pragmatic interpretation of sentences as well as the polysemy of individual words. All suspicious activity involving language is coded in some way. No terrorists, fraudsters, or paedophiles are going to openly declare their intentions in plain language. Rather, their meanings are expressed indirectly and the overall import builds up over a period of time. Any individual sentence, viewed in isolation, appears to be innocent. Only when viewed as part of a sequence with other sentences does a picture emerge of a hidden intention. In the case of paedophile activity, for example, the sentence *How old are you?* is innocent enough as a casual enquiry; but seen along with such other sentences as *Are you alone?* or *What are you wearing?*, a different linguistic profile appears. We need to distinguish between innocent conversations and those which, through their use of suggestive words and sentences, build up a suspicious pattern of discourse over time. Such analyses are not easy to make, however, for reasons that are nothing to do with linguistics. It is difficult to obtain samples of authentic data to analyse in order to provide norms. Applied linguists need to obtain clearance from the relevant authorities whenever they propose to engage in counter-criminal research, and this is never easy to obtain.

Directions for research

The above examples suggest several directions for future searchlinguistic research.

A semantic approach is needed to describe the polysemy of lexical items, relate the senses to a knowledge hierarchy representing online content, and assess the contribution individual lexical items make to the semantic identity of an e-text. It is not enough to say that *depression* has the four meanings noted above; we must assign each use to the knowledge domain to which it belongs. *Depression*='downturn in economic growth' needs to be assigned to 'economics'; *depression*='area of low pressure' to 'meteorology'; and so on. The amount of ambiguity presented by a lexical item also needs to be taken into account: an item such as *quarterback* makes a high-value contribution to the domain of 'American football', because it is rarely encountered outside this setting; when we see that word on a page, it is virtually certain that the page is going to be about American football. By contrast, an item such as *depression* is less predictive because it is used in four domains; and an item such as *country* has a very low predictive rating, because it can be used in relation to hundreds of domains (all the countries of the world, for a start). Once lexical items have been identified in this way, it is possible to build up lexical sets which specify the semantic content of a knowledge domain, and this can then be used to build a filter which will analyse the content of web pages, as in the iSense and Sitescreen products developed by an e-advertising network (ad pepper media, 2008)). These sets need to include high-frequency collocations, and they also extend the traditional remit of semantics, as proper names have to be included, in view of the fact that web content contains many brand names, models, company names, logos, and other 'encyclopedic' expressions.

A grammatical analysis needs to be made to ensure that all morphological and syntactic factors are taken into account, such as inflectional variants (*mobile phone(s)*) and compounding alternatives (*cellphone, cell-phone, cell phone*). Word-class tagging may be needed to distinguish homographs (*bear* verb vs *bear* noun). And syntactic information is required whenever we need to refer to sentence structure in order to resolve an ambiguity. In advertising, for example, clients often want their ads to appear only on pages which say 'nice' things about their product; they do not want their ads to appear on pages which say 'nasty' things. To identify these different kinds of sentiment, lists of positive and negative lexemes must be compiled. Positive items in English would include *fantastic*, *best*, *wonderful*, *marvellous*; negative items *awful*, *terrible*, *disaster, bad*. But, in any review, we have to allow for the effects of negation: *her latest recording is by no means bad*; *his new book is not one of his best*. The positive meaning of a lexical item can be reversed by the syntactic context in which it occurs. A grammatical perspective is thus critical.

A sociolinguistic or stylistic analysis perspective is needed to ensure that lexical lists are truly comprehensive - for example, including formal and informal variants (*television / telly,* and the various kinds of slang), regional differences (e.g. American vs English, such as *color / colour*, car *boot / trunk*), and within-region alternatives in spelling, punctuation, and capitalization (*judgment / judgement, Bible / bible*). Stylistic issues also arise in relation to pages with figurative or rhetorical content, such as metaphor, irony, sarcasm, and other forms of expression where the language operates at different levels. We do not want a football report which happens to refer to players 'sliding all over the pitch like ducks on an icy pond' to be classified as a page about ornithology. In this respect, poetry is likely to be the most difficult genre for a searchlinguistics to handle.

A pragmatic perspective is needed to take into the account differences which arise out of the purpose of web pages. Why does someone write a web page? Is it to inform, to entertain, to persuade, to express extreme views, to titillate, to sell... The pragmatic purpose inevitably affects the linguistic character of the page: for example, pages intended to sell products will have their own lexical character (*your account, basket, special offers*...), as well as a distinctive graphical and functional layout, and pages with extreme views will typically contain a great deal of taboo language. More fundamentally, the digital medium is changing our notions of how texts are created, as illustrated by the multi-authoring of wiki-pages. Ineed, the traditional notion of written text as having stability and determinate boundaries is called into question by sites (such as forums and chatroom interactions) where what we see can alter from one moment to the next (Crystal, 2009).

Searchlinguistics can never be a purely synchronic study, because the linguistic content of the internet is time-sensitive. Each page is time-stamped, even though the date at which a page was brought into being is often not immediately evident. Searches give the appearance of being synchronic, though in fact they present simultaneously hits from different time-periods. Disentangling the conflicts in the data (e.g. when a series of search results gives different population estimates for a country) is not always easy, and this problem is going to increase as the internet archive grows.

A diachronic perspective is also essential in resolving the ambiguities mentioned above. New terms are constantly being introduced into a language, and they have to be added to the lexical sets - for example, in 1999 any set of lexical items relating to Iraq would not have included the phrase *weapons of mass destruction* - something which became necessary in 2003. Or, to take a more commercial example, as new models of motor-car come on the market, their names and model designations have to be incorporated. The diachronic perspective also applies in retrospect. As more historical material becomes searchable, lexical sets devised for the present-day need to be adapted to be appropriate to the earlier period. The lexical set for road vehicles devised for the 2000s, for example, is not going to work well when applied to a newspaper corpus relating to mid-Victorian English, with its *broughams*, *phaetons*, and *landaus*.

Searchlinguistics is in its infancy because the industry that it aims to service is in its infancy. It is an industry, moreover, which is rapidly evolving, as new technologies become available and present fresh challenges and opportunities. A complication is that this is a highly competitive industry which is continually 'moving the goalposts', as businesses strive to keep ahead of each other, so that it is hard for applied linguists to keep up (Crystal, 2008). For example, whatever strategies might work for internet search on computers, these have to be adapted as internet access becomes increasingly routine via mobile phones. And fresh search problems are generated as the amount of text in a file reduces, as in the formal constraints imposed by Twitter (limited to 140 characters) or text-messaging (160 characters), or on internet pages where the content is predominantly visual, and analysts need to refer to the underlying metadata to obtain sufficient text to carry out any linguistic analysis at all. The way the data is coded also varies, so that it is not always easy to guarantee that a 'scrape' of a page to extract relevant text will exclude unwanted material, such as search menus, help files, and programming formulae. For such reasons, searchlinguistics presents applied linguists with some of their most intriguing challenges.

References

ad pepper media. 2008. The iSense semantic network. <http://www.isense.net>

ad pepper media. 2008. Sitescreen brand protection. <http://www.sitescreen.com>

Crystal, D. 2008. Who pays the piper calls the tune: changing linguistic goals in the service of industry. A case study. In D. Prys & B. Williams (Eds.), *Global understanding in multilingual, multimodal and multimedia Contexts* (pp. 39-46). Bangor University: Language Technologies Unit.

Crystal, D. 2009. The changing nature of text: a linguistic perspective. In E. Thoutenhoofd, W. van Peursen & A. van der Weel (Eds.), Text comparison and digital creativity. Leiden" Brill.