Natural Language Engineering **12** (4): 391–395.
doi:10.1017/S1351324906004414 Printed in the United Kingdom

Industry Watch

ROBERT DALE
Centre for Language Technology, Macquarie University, Sydney, Australia

(*Received 9 October 2006*)

The last six months have seen a number of interesting announcements and products in the commercial language technology arena. Here are some that caught my eye, and which you might find interesting too.

First, let's look at the text analytics world (more on that term below). A number of the major vendors in this space have announced new versions of their software, and a lot of the activity here appears to involve moves to ensure UIMA-compliance (remember, no one ever got fired for buying IBM):[1] Attensity has released its Attensity 4 Text Analytics Suite, which they claim is the market's first complete offering of its kind, incorporating 'new methods of searching, querying, charting, and graphing freeform text dynamically in an easy-to-use, browser-based interface', along with the obligatory UIMA annotators;[2] TEMIS has announced Luxid, its next generation Information Intelligence solution, 'designed from the ground up on UIMA';[3] and Nstein has launched a collection of 12 UIMA annotators for search and sentiment analysis.[4] But there is life beyond Fear, Uncertainty and Doubt,[5] with a number of vendors offering some quite interesting advances. ClearForest has released its Semantic Web Services product, which makes the company's text extraction and event detection capabilities available as a standard web service: you submit a text to the web service, and the ClearForest technology identifies the people, organizations, and geographical entities in the text, returning the results as XML or as a formatted web page; you can try it out at `http://sws.clearforest.com`. Inspired by the Google Maps mashup phenomenon, ClearForest has announced a Mashup Contest to see who can come up with the neatest app built on this technology.[6]

More interestingly, the component technologies that the text analytics companies produce are finding their way into mass-market customer-facing applications that

[1] See `www.research.ibm.com/UIMA` for information on IBM's Unstructured Information Management Architecture, an open-source, industrial strength platform for text analytics.
[2] `www.attensity.com`
[3] `www.temis-group.com`
[4] `www.nstein.com`
[5] A term from prehistory; see `http://en.wikipedia.org/wiki/FUD`. I'm not suggesting in the slightest that the promoters of UIMA are employing this strategy, but they might as well be, given the way vendors seem to be falling over themselves to ensure UIMA compatibility.
[6] There are some monetary prizes, with a first prize of $2000; but if you're short of cash you might be better to look at NetFlix's $1m prize for improving recommender technology: see `www.netflixprize.com`.

demonstrate their potential. Something I expect we're going to see more of is faceted search: this involves clustering the results from a search engine so that the user can progressively narrow the focus of what they are looking for, in a way that combines the more traditional alternatives of browsing a taxonomy versus free-text search. We've had faceted search in a number of online venues for a while (try out the demos at the websites of some of the earliest players in this space, Northern Light[7] and Vivisimo[8]), but inevitably it is the entry of Google and Microsoft into this space that will really shake things up: there's a lot of speculation on the net that Microsoft's 'Search Result Clustering', which you can explore at `http://rwsm.directtaps.net`, will find its way into a Microsoft product near you before very long.

Where this probably gets more interesting for the audience of this journal is when information extraction techniques are merged with clustering techniques. This year, Inform Technologies launched its Inform Publisher Services platform, signing up a number of media customers including *The New York Sun*, *The Washington Post*, and *Newsweek Interactive*. This technology, which is intended to make news sites more sticky so that visitors don't go searching for related information elsewhere on the net, leverages the site's archived content by providing clustered related documents, organized by topics and a variety of named entity types. You can get a good demonstration of the capability at Inform's own website at `www.inform.com`: for any of the displayed stories, click on the 'related subjects' link; or click on any of the story titles at `www.newsOK.com` and explore the 'Related Searches' menu that appears on the left-hand side of the page. You don't even have to leave your desktop to explore this kind of technology fusion: if you're a Google Desktop user, you can download Inxight's Search Extender for Google Desktop, which automatically groups search results by key mentions of people, companies, places and concepts;[9] Figure 1 shows an example.

Moving away a little from the core applications in this space, it looks like there is still life in the grammar-checking market. Early grammar checkers were driven by a Strunk-and-White style 'keep it simple' philosophy, eschewing verbiage and complicated language. Whitesmoke's grammar checker[10] goes in the other direction: it 'takes your writing to a new level!' by suggesting adjectives and adverbs to be added to your text. So, for example, 'I believe we can offer him a discount' becomes 'I believe we can offer him a *generous* discount'. The sofware costs $80 for the basic version, and $100 for more specialised editions, including versions targetted at the medical and legal communities, a creative writing version, and, appropriate to the Internet age, a dating version: 'WhiteSmoke will help you appear more sophisticated and appealing for your potential dates.' So maybe that seemingly articulate person who mailed you from the dating site last night is not really as eloquent as you thought. On a more serious note, the Criterion Online Writing Evaluation Service from the Educational Testing Service is worth a look: take the

[7] `www.northernlight.com`
[8] `www.vivisimo.com`
[9] `www.inxight.com/products/se_google`
[10] `www.whitesmoke.com`

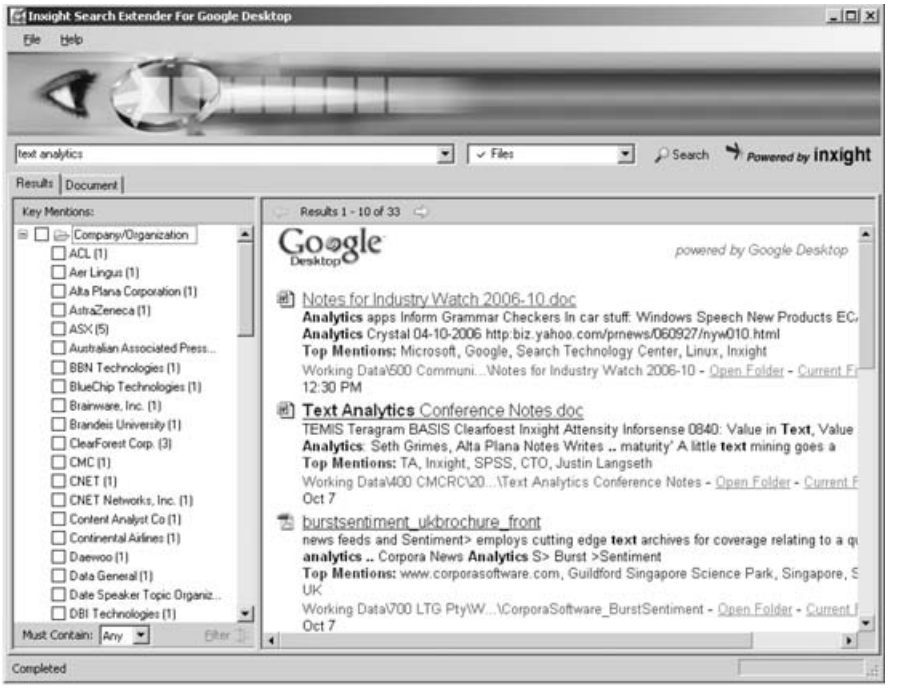

Fig. 1. Inxight's Search Extender for Google Desktop

tour at `www.ets.org/criteriontour.html`. This tool provides help with grammar and spelling, as you might expect, but the organisation and development analysis in particular looks interesting; this uses a voting algorithm whereby three separate discourse analysis systems label the elements in student essays to give some idea of the document's structure. You can find out more about how this works in Burstein et al (2003).

There have been interesting developments in the speech world too. The media continues to report conflicting sentiments on the state of the market for automatic speech recognition (ASR); not a week goes by without a news story somewhere that starts with a line like 'After 10 years in the darkness, the speech recognition market is set to explode . . . ', while there are equally many reports from analysts saying that the technology is still not mature enough to be allowed out to play on its own. In August, Fluency[11] released its *2006 Consumer Perceptions into Speech Recognition Survey*, which tells us that people are happier with speech than they were before: 69% think speech recognition systems have improved over the last year, although 22% are undecided; and a higher number of younger users (92%) think it has improved. At the same time, both Microsoft and Nuance have said that they will work with the GetHuman project,[12] which has established a database of unpublished phone numbers and codes that enable callers to skip directly to

[11] `www.fluencyvoice.com`
[12] `www.gethuman.com`

a human when they call customer service.[13] I suppose this acknowledges that, for some people, there's no ASR like no ASR.

Fluency's survey is concerned with telephony-based systems, but vendors of desktop recognition are talking things up too: Dragon's new Version 9.0 of NaturallySpeaking claims 99% accuracy out of the box without training (well, for *some* speakers, for the American English version). And Dragon claim not to be worried by the imminent release of Microsoft Vista with integrated speech recognition.[14] Meanwhile, Tesco, the UK supermarket giant, is branching out into its own software line, which doesn't need to be as expensive as Microsoft's offerings because 'we kept out stuff people just don't need like voice recognition.'[15] Again, for some people there's no ASR like no ASR ...

Still, some existing voice apps are becoming more widespread, and new ones are becoming available. A fair number of companies now offer services that turn your voice mail messages into text, sending them straight to your mobile phone or email. For UK users, SpinVox offers a free trial for one week or 50 messages;[16] SimulScribe's service in the US costs 25 cents a message, and offers a free 1 month trial.[17] There are lots of similar services available; as always, Google is your friend.

An exciting development is podcast transcription: services like PodZinger[18] (a spin-out from BBN) translate audio content into text and then provide standard search interfaces over that text, allowing rapid jumping to the audio portions you're interested in. Pluggd's interface is particularly nice:[19] given a search term, it also looks for related terms in the file, and indicates relevance on a file timeline using a heat map metaphor, so that the parts of the file most relevant to your query appear in red, less related in green and unrelated in blue. You hover over any of the marked points to see what relevant terms Pluggd found there. Meanwhile, Wizzard is also offering a service whereby text blogs will be turned automatically into podcasts.[20] I see a rash of round-trip ASR-to-TTS-and-back-again experiments on the horizon.

Talking of heated language: Nintendo has a patent application for voice recognition technology to be incorporated into multiplayer games.[21] When a user speaks, the utterance is converted into on-screen text and sent to the other players. But not just plain, boring old text: the system also picks up the tone, pitch and volume of the voice and translates that into colour, font size and character set. So, if you shout 'I win!' loudly into the microphone, what you may get on screen is allcaps text in a large red font; speak quietly and you may get a pastel shade in a smaller typeface.

But if what you are really interested in is mass-market end-user impact from speech recognition, watch out for Daewoo's new microwave oven, due for release

[13] See `http://biz.yahoo.com/prnews/060808/sftu036.html`.
[14] See `http://www.theinquirer.net/default.aspx?article=34072`.
[15] See `http://businessweek.com/globalbiz/content/oct2006/gb20061004_001766.htm`.
[16] `www.spinvox.com`
[17] `http://simulscribe.com`
[18] `www.podzinger.com`
[19] `www.pluggd.com/`
[20] See `www.wizzardsoftware.com/pr/show_news2.php`.
[21] United States Patent Application 20060025214.

in 2007: this understands 40 different spoken commands, like 'heat on high for 30 seconds' and 'cook popcorn.'[22] We've been joking about it for years and now it's finally happening.

At the start of this article I used the term 'text analytics'. I haven't checked carefully, but this might be the first time that term has appeared in the pages of this journal. A couple of years ago, the most widely-used industry term for the collection of text processing techniques that include information extraction, text summarization, clustering and classification was 'text mining'. In a change of name that is just a little reminiscent of the rebadging rife in the speech recognition industry (where this year's label is 'voice self service'), 'text mining' seems to have been dropped in favour of 'text analytics.' This shift from blue-collar to white-collar connotation was cemented by the 2nd Annual Text Analytics Summit, held in Boston in June 2006: no, you didn't miss the 1st Annual Text Analytics Summit; the preceding event in the series was called the Text Mining Summit 2005.

At the time of writing, Google found 1,640,000 hits for text mining but only 101,000 for text analytics, but I predict that will look very different within a year, unless of course some other term ('information access' and 'content access' seem to be gaining a little purchase) appears more effective in communicating to the rest of the world what this stuff is really all about.

Reference

Burstein, J., Marcu, D., and Knight, K. (2003) Finding the WRITE stuff: Automatic identification of discourse structure in student essays. In S. Harabagiu and F. Ciravegna (Eds.), *IEEE Intelligent Systems: Special Issue on Advances in Natural Language Processing*, **18**(1): 32–39.

[22] See `http://www.slashgear.com/voice-recognition-microwave-from-daewoo-271850.php`.