Natural Language Engineering **13** (3): 283–286. © 2007 Cambridge University Press
doi:10.1017/S1351324907004603 Printed in the United Kingdom

*Industry Watch**

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

(*Received 10 July 2007*)

I've just come back from the 45th Annual Meeting of the Association for Computational Linguistics (ACL) in Prague; this was the biggest ever ACL conference, with more than 1,000 people attending for the first time. Attendance at ACL conferences has been growing year on year, and that is a sign of a healthy field. Another sign of health is industry sponsorship. For this year's conference, the Gold Sponsor was Google, and Microsoft and Yahoo! were Silver Sponsors, along with a few companies we have not seen as ACL sponsors before: _textkernel,[1] NewsTin,[2] and – a name that seems now to pop up regularly in this column – Powerset.[3] There are all sorts of reasons why companies sponsor conferences like this, but clearly a major purpose is to make themselves visible to potential employees. And, if companies are hiring, that's good news across the board: it gives us a way of attracting more students into the field, and more generally, it speaks to the industrial and commercial relevance of what we do. There is nothing like external validation – especially commercial validation – to wash away those niggling self-doubts about the utility of your research endeavours. I remember attending MT Summit VII in Singapore in 1999, when Jo Lernout, then of Lernout & Hauspie, gave an invited talk in which (I'm sure I'm remembering this correctly) he said his ambition was to hire everyone in the hall. There were around 250 attendees – big for an NLP conference at the time – so that created quite a buzz. Jo wanted to hire *everyone*, not just cherry pick those with the near-to-product big ideas; in his vision of the future, every teensy-weensy tightly-focused research contribution had a role to play. For just a moment, everybody felt wanted.

Anyway, back to Powerset. Not only were they a major sponsor at the ACL conference; Barney Pell, Powerset's CEO, also gave an invited talk, entitled 'Powerset and Natural Language Search'. Powerset, as most of you will by now surely know, is the until-recently-in-stealth-mode company that aims to turn the search world upside down by providing the ability to pose natural language questions to the Web, via its

* Industry Watch is a semiregular column that looks at commercial applications of natural language technology. The author can be contacted at `rdale@acm.org`.

[1] `www.textkernel.com.`

[2] `www.newstin.com.`

[3] `www.powerset.com.`

https://doi.org/10.1017/S1351324907004603 Published online by Cambridge University Press

exclusive licence to use Xerox's XLE technology.[4] Barney's presentation generated mixed reactions: a number of people I talked to felt it was a little off-target, with the first half of the talk being too much like a sales pitch. Barney did start to get to technical aspects in the second half of the talk, but this is a conference whose attendees are pretty much all NL-heads, more interested in algorithms than in advertising; he was preaching to the converted, who don't need to be convinced that natural language is the future of search. Hey, we believe this almost as a matter of faith.

Speaking of faith: funnily enough, my bedtime reading during the conference was Richard Dawkins' *The God Delusion*. Irrespective of what you think of Dawkins' 'militant atheist' stance, he makes some very interesting observations about the forces that drive religious belief. The book was very much in my mind as I sat there with the assembled masses – around 1,000 disciples of the Church of Latter Day Search – listening to Barney's sermon.[5] There is something very interesting about the psychology of all of this. As a community, we want Powerset to succeed Google-scale, because that will vindicate our way of life and our deepest held beliefs about the search for meaning. And, Powerset is ready to take its disciples, the true believers, on a mission: by the time you read this, if all has gone to schedule, Powerlabs will be open for business in true Web 2.0 style, leveraging community effort. Here's Mark Johnson, Powerset's Product Manager for Powerlabs[6]:

> Being a member of Powerlabs when we open is going to be just like being on the product management team here at Powerset because you'll be able to . . . run searches on the Powerset engine and see what our cool capabilities are, and you'll also be able to give feedback on the results which will help to train Powerset and change the way results come back in the future . . . . Most importantly you'll be able to contribute your own ideas; tell us what's wrong with search today and how we can fix it.

And here is Steve Newcomb, a Powerset cofounder and COO:[7]

> We want as many people in Powerlabs to help us build and test the product [sic]. Powerlabs tells us when we are ready to go. We could have 50,000 people QAing our product.

Pondering all this, I couldn't help but see a parallel between Barney's invited talk and Jo Lernout's of several years before. Both painted a picture where what we all did day-to-day was important, and made sense in a bigger picture that would change the world; that makes people feel good. And like Jo, Barney wanted to hire everyone in the hall, except that he's not paying,[8] and he doesn't have to; we'll contribute willingly because we *believe*.

[4] See the *Industry Watch* columns in the previous two issues of this journal for some discussion of Powerset. There is information about XLE at `http://www2.parc.com/isl/groups/nltt/xle/`.

[5] These thoughts were possibly triggered by my sudden realisation that Barney shares his surname with Cardinal George Pell, the Catholic Archbishop of Sydney.

[6] This is transcribed from Mark's video on You Tube at `www.youtube.com/watch?v=8D6czWVYc-o`.

[7] See `http://blogs.zdnet.com/BTL/?p=5541`. That same report tells us that community participants will have to be certified for Powerlabs, and suggests this means that individuals have to go through some training to be of help to Powerset.

[8] As it turned out, neither, of course, was Jo: L&H went bankrupt in October 2001 following accusations of malpractice and overstated earnings, leading to court proceedings. The trial started in Ghent in May this year. See `www.boston.com/news/local/massachusetts/articles/2007/05/21/lernout__hauspie_fraud_trial_opens/`.

https://doi.org/10.1017/S1351324907004603 Published online by Cambridge University Press

There's something about this that makes me just a bit uneasy. At the time of writing, only a chosen few have so far seen Powerset's demo. Of course, Powerset is hardly the only QA-kid on the block, and I expect the guys at the other places must be more than a little annoyed at the way Powerset is working this. For example, CognitionSearch[9] is being used by a real customer, and Hakia[10] is up on the Web for all to play with, but they get next to zero media and blog coverage in comparison to Powerset. They are clearly not being evangelical enough. Or maybe it's easier to believe in gods you can't see.

Speaking of QA systems, a new one I came across on trawling the Web for this installment of the column is Dialogus.ru.[11] The first time I used this, the Web site displayed some cute dynamic graphics that showed all the stages of linguistic processing that were presumably being invoked to answer my question. It was nice to know that part-of-speech tagging, parsing, semantic analysis and pragmatic analysis were all coming into play, but showing these graphics meant that it took about 20 seconds for the answer to my question to appear. On a more recent visit to the site, I notice that the graphics have now gone, resulting in a faster delivery of answers (so presumably they were not just hiding processing delays). I asked the system, *What technology does dialogus use?*, and got the rather cryptic answer:

> You too are writing still in the plain style of the honest man about esoteric speech, but you are not practicing it. What is your thought on that?

Is this just the QA-equivalent of speaking in tongues, or is it an instance of statistical machine translation applied to a Bible verse in Russian?[12]

The last few months have seen a bit of realignment in the text mining space. Business intelligence company Business Objects announced that it was acquiring Inxight Software,[13] and Reuters announced plans to acquire ClearForest to help manage large amounts of financial information.[14] As of the time of writing, Inxight's Search Extender for Google Desktop and ClearForest's Gnosis plug-in for Firefox are both still freely available, but these are just the kinds of things that could disappear as a consequence of changes in strategy following acquisition, so it might be a good idea to visit the sites and grab your copies now.[15] Reuters is also making use of sentiment analysis developed by Corpora:[16] they have a new market data

[9] `http://cognitionsearch.com`.
[10] `www.hakia.com`.
[11] `http://dialogus.ru`.
[12] A more recent visit to the site suggests that it might just be a front for a porno site: the sample question I was invited to try was 'Do you have pictures of naked women?'
[13] `http://news.com.com/2110-1012_3-6185692.html`.
[14] `www.econtentmag.com/Articles/ArticleReader.aspx?ArticleID=36041`.
[15] Inxight's Search Extender automatically groups search results by key mentions of people, companies, places and concepts; see `www.inxight.com/products/se google`. Gnosis carries out named entity recognition on the Web page you are reading; see `http://gnosis.clearforest.com`.
[16] `www.corporasoftware.com`.

https://doi.org/10.1017/S1351324907004603 Published online by Cambridge University Press

product that scans company news articles and analyses the sentiment they contain, using this information to trigger orders in algorithmic trading applications.[17]

Returning to an earlier theme: of course, Powerset is not the first company to leverage an army of Web users to improve its product, and it certainly won't be the last. Many have wondered what business model lies behind Google's free 411 service,[18] which uses voice recognition to provide business listings over the phone. Tim O'Reilly speculates[19] that it's really all about gathering vast amounts of speech data for future products. If you haven't tried the service, or don't live in the United States, you can get an idea of what it's like at `www.webpronews.com/blogtalk/2007/05/24/experiment-with-goog-411`.

And here's one for the 'if you can't beat them, join them' collection. Voice recognition 'self-service' systems and call centers in developing countries have become competing solutions in the battle to provide ever-cheaper customer service in a world of shrinking margins. Each has its own problems, of course; voice recognition systems are prone to error, but it can also be hard to understand the heavy accent of someone who speaks a different brand of English to your own. Carnegie Speech[20] is coming to the rescue: Carnegie Speech Assessment is software that listens to a user's speech to determine what's correct or incorrect in the sound, rhythm and pitch, so that it can provide an overall assessment of the quality of the user's speech; and the company's NativeAccent product pinpoints a user's pronunciation errors and then shows the user how to correct them. It's only a matter of time before someone runs the software on the output of a text-to-speech system.

Finally, this quarter's entry for the cute applications department: Xerox is developing technology whose purpose is to make the adjustment of colours in a document as easy as simply describing the colour you want. Using either keyboard or voice, users can say things like *make the sky a deeper blue* or *make the background carnation pink*, and the software makes the appropriate changes. The work is still at the research stage;[21] you can find a paper that talks more about the idea at `www.xerox.com/innovation/simple_color.pdf`.

[17] `www.finextra.com/fullstory.asp?id=16869`.
[18] `http://labs.google.com/goog411/`.
[19] `http://radar.oreilly.com/archives/2007/04/why_google_is_o.html`.
[20] `www.carnegiespeech.com`.
[21] `www.webwire.com/ViewPressRel.asp?aId=34340`.

https://doi.org/10.1017/S1351324907004603 Published online by Cambridge University Press