Natural Language Engineering **11** (1): 1–5. © 2004 Cambridge University Press
DOI: 10.1017/S1351324904003663 Printed in the United Kingdom

Industry Watch

ROBERT DALE
Centre for Language Technology Macquarie University Sydney, Australia

(*Received 15 November 2004*)

One way to keep in touch with what is happening in the commercial speech and language technology world is to pay occasional visits to the websites of HLT Central (at `www.hltcentral.org`) and LT World (at `www.lt-world.org`). Both sites provide links to news stories and press releases from companies and other organizations active in the area. The people who run these sites trawl the web for news stories of relevance, saving you the trouble of doing that yourself.[1]

The stories come from a variety of sources: the HLT Central team trawl for news stories from the web, on a daily basis, with some basic keyword combinations; and they monitor the news pages of about 120 language technology companies worldwide, as well as receiving material directly from a small number of companies. Since there's a human in the loop here, you can expect the material provided on the site to be relevant to the topic; in other words, precision is good. But inevitably there's great potential for weakness in recall: there could be language technology news hiding in places where one might not think to look, a point we'll return to below.

Despite this shortcoming, if you're interested in the commercialisation of speech and language technologies, the information collated on this site provides good insights into what is happening out there in the real world. For the period from January 2004 to early November 2004, the time when this column was written, HLT Central lists 653 articles, an average of about three per working day. I decided to take a look at these to see if they would provide some interesting observations on the current state of the speech and language technology industry, and of recent developments therein.

Of these 653 articles, a quick scan suggested that 68 were general interest pieces of one sort or another: for example, there are news stories about the pros and cons of spam filtering and speech-to-speech translation, from sources such as *BBC News* and *Wired*. Although articles of this kind can provide important data points in terms of the public perception of these technologies, what I was more particularly interested in was concrete commercial activity in terms of products and services, so we'll exclude the general interest stories from further discussion.

I took the remaining 585 stories that appeared to be specific technical announcements, and separated these out into those concerned with speech technology, and

[1] At the time of writing, HLT Central was seeking sponsors for the continued operation of the web site. The site provides a wide range of immensely useful information; if you work for an organization that might be able to sponsor HLT Central's activities, please contact `arax@HLTCentral.org`.

those concerned with text-based technologies. This is not hard to do, since the HLT Central site already provides some basic category labels, although inevitably there are inconsistencies in the application of the categories. I found that 513 stories—that's 87%—were concerned with speech technology, and only 72, or just under 13%, were concerned with what we might think of as text-based technologies. Now, it has to be stressed that this is just an informal count; in particular, because of the rough and ready way in which I did this, I probably missed a number of general interest pieces amongst the stories on speech. But even allowing for this, the difference in number of articles in the two broad areas is striking.

Given the above-mentioned problem with recall in terms of the data gathering exercise represented here, this can't be taken as a reliable measure of the relative strengths of the speech and text-based sectors of the industry. There could be all sorts of reasons as to why speech might be more visible in the marketplace, but we'll leave that for another column. In this column, I want to look in particular at what text-based technologies were making news in 2004. Here are some observations on what is talked about in the 72 stories that were about products and services in this area.

Machine translation

The single largest cluster of stories is concerned with machine translation (MT) or localisation. I've put these in one group because the boundary between the two can be fuzzy: I classified 21 stories as being about MT and 15 as being about localisation, but others might get slightly different numbers. Most readers of this journal would not consider many of the available localisation products or services to be language technology, but some of the more sophisticated localisation technologies embody elements of MT; and, on the other side, some of the MT products are translation-memory tools which some language technologists might not want to think of as MT.

Some of the announcements here are just product upgrades from the usual suspects, Systran (`www.systransoft.com`), TRADOS (`www.trados.com`) and SDL International (`www.sdl.com`); but there were two things that really stuck out in a scan over the MT announcements.

First, Language Weaver (`www.languageweaver.com`), the statistical MT company spun out of ISI by Kevin Knight and Daniel Marcu, rated a mention in eight stories (a substantial proportion of which were of course press releases from the company); reading these, it really does look like statistical MT is beginning to make an impact outside the research laboratory.

Second, we're beginning to see product offerings that combine MT with speech. Commercialisation of the PDA-based speech to speech translation advances made in the DARPA Babylon project is already with us: VoxTec (`www.voxtec.com`) are taking orders for their PhraseLator™ handheld device, which covers 15000 phrases in 53 languages by switching removable media. I couldn't see any pricing information on their web site; but the device looks a bit brick-like, and the available plug-in language modules available so far are tailored to the military and law enforcement sectors, so I think we can assume this product is not currently targetted at the general

population. On the other hand, NEC (www.nec.com) have produced a PDA-based Japanese-English translator, incorporating speech recognition, machine translation and speech synthesis, with a vocabulary of 50,000 Japanese and 25,000 English travel and tourism related words. This hasn't, as far as I can tell, actually hit the market yet, but apparently it was trialled at Narita Airport earlier this year as a communication aid for shop assistants talking to English speaking tourists.

If you have your doubts about the reliability of speech recognition on such devices, you could just spend US$59.95 on Franklin's 5-Language Commununicator™, a consumer-grade text-to-speech translation device that featured in the news. This is a smaller-than-PDA device that lets you select one of 6000 phrases on a small screen; it then translates and speaks your selected phrase in English, French, German, Italian or Spanish. Franklin do a number of simple but useful handheld translation devices; check out the 'Travel' section at www.franklin.com. These are all built around straightforward word and phrase lookup, which you might not want to call MT; but they are an interesting and useful step forward from phrase books.

Text search and question answering

After the MT and Localisation cluster, the next biggest category of stories collects together text search, question answering, and text mining, with a total of 23 articles.

A number of stories are about reasonably established companies in the document processing space whose bread and butter still seems to be term-based information retrieval and document categorisation. Companies like Convera (www.convera.com) and Inxight (www.inxight.com) produce tool suites in these areas, and advertise more sophisticated features that appear to take advantage of natural language processing techniques as icing on the cake. So, for example, Convera offers concept-based search using what it calls Semantic Network Cartridges and Taxonomy Cartridges; and Inxight's SmartDiscovery™ makes use of named entity recognition. TEMIS (www.temis-group.com), who acquired Xerox Corporation's linguistic product operations in 2003, offer a range of content management tools that appear to be based on reasonably sophisticated language processing. In all these cases it's hard to know what's underneath the covers, of course; but for these established players, the most obvious thing to note is that the smart capabilities are just one part of a much broader range of functionalities. Stories about these companies tend to be about new contracts or partnerships; there's not much evidence of new advances in their underlying technologies, although of course that doesn't mean such advances are not taking place.

The second subcategory is concerned with question answering. The line between conventional term-based text search and more sophisticated question answering has been blurred for a while; search engine providers don't tell you what kind of processing they apply to queries, and while it seems unlikely that they are using anything like the techniques reported on in the context of the TREC QA track, for example, it's also not clear how much sophistication you need before you can say you do question answering.

So, just as the speech recognition industry has in the last few years started talking about 'natural language speech recognition', providers of search technology now talk about suporting 'natural language queries'. ISYS (www.isysusa.com), for example, has an online demo where you can try out this functionality. I suspect this particular example does not much more than strip out words that are deemed irrelevant for searching, but there's clearly interest in doing more sophisticated query processing: Kanisa (www.kanisa.com), for example, has technology which they claim interprets the intent of a user's question 'by combining complex concept understanding with dictionaries and business-specific knowledge'.

Other vendors focus more on question-answering capabilities as their unique selling point. AskJeeves (www.ask.com) has obviously been active on this front for a long time, and another familiar name is iPhrase (www.iphrase.com), who promise a better search experience by using natural language processing on both queries and documents. Again, news from these companies tends to be about new contracts rather than new products.

Where there does seem to be something new happening is in the integration of QA technology, in whatever form, with other modes of delivery. So, Linguit GmbH's (www.linguit.com) Nuggets product lets you send a question via a text message; it finds an answer on the web, then returns the answer to your phone. Answers Anywhere™ from iAnywhere Solutions (www.ianywhere.com) offers question answering via speech, web query, or text message, by translating natural language queries into the relevant commands for back-end systems to process.

Last but not least, Connexor (www.connexor.com) deserves a mention for being the text technology company with the largest number of stories, with seven press releases. In the news reviewed for this year, this is the only company that is offering basic component technology in the form of parsers and morphological analysers, as opposed to more complete solutions.

Text entry

The final cluster of articles is about text input: five stories are concerned with text entry either on small devices or in languages where text input is an issue. Of these, four are about Zi Corporation (www.zicorp.com), who produce predictive text entry technology for small devices, with their eZiTap™ and eZiText™ products; the fifth is about Decuma AB's (www.decuma.com) handwriting recognition software for PDAs in a variety of languages including Japanese and Chinese. It's sobering to think that the first handheld device supporting handwriting, the Apple Newton, was released over 10 years ago; the technology has advanced somewhat in that time, but it's also been supplanted to a considerable extent by the apparent ease with which people can type with two thumbs.

So what does this snapshot of industry activity tell us? A number of things struck me on going through these articles.

One is the extent to which new products were largely the result of the combination of existing language technologies with specific forms of delivery, and in particular the incorporation of language technologies—machine translation, question answering, and language modelling for text prediction—into handheld devices, whether they be PDAs or phones. New platforms offer new opportunities.

In contrast, there seemed to be very few announcements that heralded breakthroughs in the underlying language technology itself; the only really obvious example of a language technology moving from research into commercial use during the period surveyed, as noted above, was Language Weaver's statistical MT. No doubt there are a number of other incremental technical advances underlying some of the announcements, but these don't make news.

The more important point here is that the market is not interested in natural language processing *per se*, it is interested in solutions. How these solutions are arrived at really doesn't matter; it's the end, not the means, that is important. One manifestation of this last point is the difficulty of determining precisely what—if any—natural language processing is being used in the products on offer. Of course, it's not in the interests of the vendors to reveal how they do what they do, and they shouldn't be expected to; but this makes it hard for the field to identify and claim successes, which has an impact on the general perception of the utility of research in LT.

I've already alluded to the fact that this data collection exercise is noisy and imprecise, and so you might be forgiven for taking any observations deriving from it with a pinch of salt. This might particularly be the case with respect to the speech vs text-based technology balance. I suspect that there are significant text-based language technology advances out there, but these are never characterised in those terms; that's part of the reason that they are not discovered by the data gathering techniques used by web sites like HLT Central and LT World.

So here's a place where a product based on cutting-edge language technology could used to promote the visibility of the field as a whole: how about a concept-based search tool that can crawl the web and produce a list of links to new product offerings that are truly in the speech and language technology areas, even when those offerings aren't characterised in terms of our conventional research-based terminology? There are some interesting language processing challenges in there that go beyond the notion of concept search being promoted by search engine vendors; and there's great potential for something like this in the business intelligence market.