Natural Language Engineering **21** (4): 653–659.

doi:10.1017/S1351324915000200

Industry Watch

NLP meets the cloud

ROBERT DALE
Chief Technology Officer, Arria NLG plc
e-mail: Robert.Dale@arria.com

(*Received 17 June 2015*)

Abstract

With NLP services now widely available via cloud APIs, tasks like named entity recognition and sentiment analysis are virtually commodities. We look at what's on offer, and make some suggestions for how to get rich.

Software as a service, or SaaS—the mode of software delivery where you pay a monthly or annual subscription to use a cloud-based service, rather than having a piece of software installed on your desktop—just gets more and more popular. If you're a user of Evernote or CrashPlan, or in fact even Gmail or Google Docs, you've used SaaS. The biggest impact of the model is in the world of enterprise software, with applications like Salesforce, Netsuite and Concur now part of the furniture for many organisations. SaaS is big business: depending on which industry analyst you trust, the SaaS market will be worth somewhere between US$70 billion and US$120 billion by 2018. The benefits from the software vendor's point of view are well known: you only have one instance of your software to maintain and upgrade, provisioning can be handled elastically, the revenue model is very attractive, and you get better control of your intellectual property. And customers like the hassle-free access from any web-enabled device without setup or maintenance, the ability to turn subscriptions on and off with no up-front licence fees, and not having to talk to the IT department to get what they want.

The SaaS model meets the NLP world in the area of cloud-based microservices: a specific form of SaaS where you deliver a small, well defined, modular set of services through some lightweight mechanism. By combining NLP microservices in novel ways with other functionalities, you can easily build a sophisticated mashup that might just net you an early retirement. The economics of commercial NLP microservices offerings make these an appealing way to get your app up and running without having to build all the bits yourself, with your costs scaling comfortably with the success of your innovation.

So what is out there in the NLP microservices space? That early retirement thing sounded good to me, so I decided to take a look. But here's the thing: I'm lazy. I want to know with minimal effort whether someone's toolset is going to do the

job for me; I don't want to spend hours digging through a website to understand what's on offer. So, I decided to evaluate SaaS offerings in the NLP space using, appropriately, the Short Attention Span (SAS) methodology: I would see how many functioning NLP service vendors I could track down in an afternoon on the web, and I would give each website a maximum of five minutes of exploration time to see what it offered up. If after five minutes on a site I couldn't really form a clear picture of what was on offer, how to use it, or what it would cost me, I would move on. Expecting me to read more than a paragraph of text is so Gen X.

Before we get into specifics, some general comments about the nature of these services are in order, because what's striking is the similarities that hold across the different providers. Taken together, these almost constitute a playbook for rolling out a SaaS offering in this space.

Demos: In terms of drawing people in, it's imperative that you have a fully functioning demonstration of the service capability right there on the site. Faced with so many vendors who are willing to let you try out the product, anyone who doesn't do this is going to have a hard time gaining any traction. And it's not enough just to have a demo: the demo also needs to be easy to find, easy to use, and it needs to produce output that's easy to understand. If you want me to believe in your product, I should be able to provide my own text for analysis as well as using samples suggested by the site.

Multiple functionalities: The days where you could offer a single niche functionality are gone, if they ever existed. Every one of the more visible and successful NLP microservice vendors offers a portfolio of services: entity extraction, sentiment analysis, classification and language identification are very common, but most vendors pad this out with a range of other functions so that they can claim 10 or so distinct APIs. It's important to provide explanations of what those functionalities are, especially for the more esoteric ones. Some sites provide video explanations of important concepts, but my SAS methodology doesn't allow time for watching videos, unless they have cats in them.

Multiple natural languages: A number of vendors provide a table that shows which functionalities they offer in which languages. Of course, this tells you nothing about the quality of performance across the different languages; from a marketing perspective, putting a tick in the relevant cell may require only the most minimal of justifications.

API flexibility: By default, everything is provided via a RESTful service, with JSON being the typical output format. To use any of these services (including free trial usage of the APIs outside of the website demos), you need to sign up for an API key. Many vendors also provide SDKs for a variety of programming languages, in an attempt to remove that last bit of friction around actually writing code to call the API. A number of vendors offer an on-premises version of their software for those clients who are uncomfortable shipping data into the cloud (still a big concern for many enterprises). A few offer a spreadsheet plug-in, so you can be up and running doing sentiment analysis on your desktop in a familiar tool in no time at all.

Visible documentation: By exposing the API documentation to public view on the website with no restrictions, the vendor shows confidence in their offering, and gives potential buyers a real sense of what it will be like to use the API.

Pricing: Almost every vendor uses a standard 'five-level' set of subscription plans, where level 1 is some number of free transactions; levels 2, 3 and 4 are package sizes that give you ever-cheaper per-transaction rates as you buy more transactions; and level 5 is the 'call us to a get a deal on more' category. Actually comparing costs here is a bit like comparing mobile phone plans, since every vendor charges different amounts and provides different numbers of daily or monthly transactions. On top of that, there are minor variations in exactly what counts as a transaction, differences in levels of support (such as how much you have to pay to transition from email to phone support), limits on the rate of transactions (i.e., number of hits per second), and other bits and bobs that make detailed comparison tricky. Overall, though, you can expect to pay something like a quarter of a cent (0.25c) per transaction on the lower-volume plans, with the price dropping to a fifth of that (so, 0.05c per transaction) for higher volumes. Some vendors also offer a pay-as-you-go model as an alternative to paying a monthly subscription, and some give a discount if you pay for a year up front.

As suggested above, there's a lot of commonality across these offerings, so I was mostly interested in what distinguished one service from another. After a few hours of exploration, five offerings had passed my SAS test. All do something reasonable in terms of the various dimensions outlined above. Here they are, in alphabetic order. Note that websites for products like this change often, so many of the specifics here may no longer hold true by the time you read this.

1 AlchemyAPI

AlchemyAPI (www.alchemyapi.com) has been around since 2005, but was acquired by IBM in early 2015. The site's key message on the main page is *Build Smarter Apps with AlchemyLanguage: 12 Semantic Text Analysis APIs Using Natural Language*. No messing around here: an unmissable link to the demo and another for the API key sign-up are right there in the middle of the page, along with a few other calls to action. There's a comprehensive set of functions that makes the site feel like a one-stop shop: Alchemy's 12 'semantic APIs' are entity extraction, sentiment analysis, keyword extraction, concept tagging, relation extraction, language detection, text extraction, microformats parsing, feed detection and linked data (whereby, when you run entity extraction, you get links to a variety of online information sources for the identified entities). Alchemy claims to support 'more than half-a-dozen languages', including English, Spanish, German, Russian and Italian.

The demo is very nicely presented: you can select from a number of suggestions for either a URL or a text fragment to process, or you can provide your own. Just for fun, I tried it on a paragraph from the last *Industry Watch* column (paragraph 3, henceforth to be known as 'the Johnny Depp test'):

https://doi.org/10.1017/S1351324915000200 Published online by Cambridge University Press

> But as future technologies go, portrayals of Artificial Intelligence are a bit different, because the reality horizon doesn't move. Human intelligence today is what it was 100 years ago. So, as time goes on, cinematic portrayals of machine intelligence, and particularly those that are manifested via the use of language, seem less and less like fiction. If you screw up your eyes a bit, and stop watching before the AI actually gets out of hand (I'm trying to avoid a spoiler here), the machine version of Johnny Depp in the 2014 movie Transcendence does seem just a wee bit like IBM's Watson on steroids.

This is hardly a carefully crafted test text, but as I said already, I'm lazy, and I had this close at hand. Although Alchemy picked up Johnny Depp as a person and linked to his DBpedia entry, it failed to recognize the reference to IBM Watson, and it didn't recognize the movie title. There were a number of other questionable results: based on the output from the samples provided by Alchemy, I initially thought the relation extraction was based on a syntactic parsing capability, but an analysis of the results of the Johnny Depp test suggests that this just uses very simple heuristics. Of course, any one-off test like this is going to produce idiosyncratic results; your mileage will definitely vary. But I suspect few potential customers are going to carry out a more rigorous scientific test than this on a first visit, and first impressions count.

Aside from these basic language APIs, and under a separate subscription scheme, Alchemy also offers a higher-level API called the AlchemyData News API that lets you interrogate a news and blogs dataset that indexes 250–300k articles every day. Broadening out from the language-specific capabilities, Alchemy has also recently added an API for a computer vision service that tags images.

2 Aylien

Aylien (www.aylien.com) is an outfit based in Dublin, Ireland. Their key landing page message is *Unlock the Hidden Value of your Text*. The site claims they offer eight APIs, but their demo actually identifies 11 capabilities: article extraction, concept extraction, entity extraction, summarisation, classification, semantic labelling, image tagging, sentiment analysis, hashtag suggestion, language detection and microformat extraction. Most of these functionalities are available in six languages: English, German, French, Italian, Spanish and Portuguese.

Again, you get the full range of options for testing via an online demo: text or URL, vendor sample or user provided. Whereas AlchemyAPI's demo interface runs all the APIs on your selected text, Aylien requires you to run it separately for each API, which I found a little frustrating. The presentation of results here was also a little confusing: you get to see the processed text with named entities highlighted *in situ* (the 'annotated view'), followed by a separate list of the entities found, but the two views didn't appear to be consistent. In the Johnny Depp test, the annotated view suggested that poor old Johnny had been passed over completely, but Johnny Depp did appear as a single entity in the results list.

Aylien's most interesting distinguishing feature is a text analysis add-on for Google Spreadsheets that lets you run the API functionalities via a familiar interface. This is a neat way of addressing the 'business analyst' audience who don't want to mess

https://doi.org/10.1017/S1351324915000200 Published online by Cambridge University Press

around with an API. Aylien also provide a 'sandbox' on the website where, provided you already have an API key, you can easily test calls to the API. Coming soon, perhaps in response to the AlchemyData API, they promise a News API that filters content from over 50 major sources.

3 Lexalytics/Semantria

Lexalytics has been around since 2003, offering sentiment analysis via its Salience engine, which is marketed as an on-premise solution. Semantria, founded in 2011, started out with the goal of making sentiment analysis accessible for US$1000 in less than three minutes, and ended up with a cloud API and Excel plug-in to address that goal. Semantria was acquired by Lexalytics in mid-2014.

Lexalytics' key message is *State-of-the-art technologies to turn unstructured text into useful data*. At this point in time, my impression is that the integration of the Lexalytics and Semantria technology bases is ongoing, resulting in a slightly muddled identity on the website: the landing page seems to position Salience and Semantria as two alternative products, but the demo links for both go to the same place. The pricing structure is somewhat different to that of the other vendors, and looks like an attempt at a compromise between a seat-based model for the Excel app and the cloud-based subscription model, with the cheapest API package being the same price as the Excel app (which makes it look expensive compared to the other providers).

Semantria's online demo allows the usual 'our text or yours' options for analysis, but is less comprehensive in its outputs. In the Johnny Depp test, it picked up Johnny but missed Watson. You have to dig a bit to get an idea of the full range of APIs available: it looks like this covers sentiment analysis, concept extraction, categorisation, named entity extraction, theme extraction, and summarisation. The site comes with what looks like a reasonable collection of tutorial videos.

Semantria natively covers 10 languages: English, French, Portuguese, Spanish, German, Mandarin, Italian, Korean, Japanese and Dutch. Other languages are available via 'paid add-ons through partners', which suggests an interesting model for growth. There's an SDK with libraries in C++, Java, PHP, NET, Python, Ruby and JavaScript.

Overall, this was the most confusing of the sites that made it into my top five. I think there's a lot of useful stuff here, but the relationship between the combined company's two products needs to be cleaned up, and there's a whole lot that the Lexalytics team could learn from the way other vendors are marketing their wares.

4 Meaning cloud

Meaning Cloud (www.meaningcloud.com) used to be called Textalytics, but went through a name change in early 2015. Meaning Cloud's landing page pitch is *Extract valuable information from any text source*. Like a number of other vendors here, you

https://doi.org/10.1017/S1351324915000200 Published online by Cambridge University Press

can use their technology via an Excel add-in or via the cloud APIs, although it looks like the full range of functionalities is only available via the latter.

Meaning Cloud offers topic extraction, text classification, sentiment analysis, language identification, linguistic analysis (lemmatisation, part of speech tagging and parsing), text proofreading and corporate reputation. There are SDKs for PHP, Java, Python and Visual Basic; there's also a plug-in for GATE. In addition, in an effort to help companies in specific verticals, the company has two other APIs: a media analysis API, which is designed to provide a high-level analysis of mentions, topics, opinions and facts; and a semantic publishing API, which combines a number of natural language processing functions that can help publishers more efficiently categorize, manage and produce content.

The demos on the site are also separated into these two verticals, but they appear to be just different ways of presenting essentially the same functionality. The semantic publishing demo identified Johnny Depp as an entity and linked to his Wikipedia page; it recognized Watson as a place.

There's what looks like a really nice set of test consoles that offer dialog box interface access to a lot of API parameters. Unfortunately, you need an API key to use these: it would have been nice to see these on this side of the key-request fence.

5 TextRazor

TextRazor fixes jetlag. I know this because I am typing this on the morning after a long-haul flight to San Francisco from Sydney. I was in a dozy state, but dozy state no more: TextRazor's bright red landing page jolted me awake in a way that a Vente Americano from Starbuck's across the road could not accomplish (I know, that's not saying much).

TextRazor's landing page message is *Extract Meaning from your Text*. The demo link is centrally positioned on the page, like Alchemy's. The demo is not as slickly presented, but the results are potentially more interesting if you have a linguistic bent: you can view an analysis in terms of words, phrases, relations, entities, meaning and a dependency parse. It's not immediately clear how you would use all of these, and the results are, as always, somewhat variable; but this was the winner on the Johnny Depp test, not only finding Johnny and linking him to his web page, but also finding Transcendence the movie.

The API offers entity extraction, disambiguation and linking, key phrase extraction, automatic topic tagging and classification in 10 languages. Official SDKs are provided for Python, PHP and Java. Overall, TextRazor transactions come in at something like half the cost of Alchemy's, but there are a number of features that also make it look more interesting: as noted above, the API includes dependency parsing, but it also supports customization, and it comes with a Prolog rules engine that lets you extend the behaviour of the API. If you already know something about NLP, this might be your best bet, since it gives you more control and flexibility than the other tools. On the downside, I had some script errors when running the online demo in Firefox, although it seemed to be ok in Chrome.

6 Summing up

So there you go: five NLP SaaS offerings as of June 2015. If yours is interestingly different from the above, I'm sorry I missed it: drop me an email.

My take on this is that the provision of NLP services via the cloud is now already quite a mature area, with a number of well-established players, and a common set of functionalities that have essentially been commoditised. We might think of this as the 'generic text analytics market'. At this point it's going to be pretty hard to be a new entrant into that market, and the existing players are going to have to distinguish themselves in some way. One way to do that would be in terms of quality, but from a technical perspective it's hard to assess the relative quality of the different offerings without doing a much more thoroughgoing test than I used here. In the short term, at least, qualitative assessments will likely be based on hearsay.

I think it's more likely that we'll see the competition distinguishing itself via different value-added functionalities like the news feed data APIs, add-ins for common desktop platforms, and end-user configuration capabilities. Each of these speaks to the needs of a particular audience, so we might expect to see some of the existing vendors specialise in a more carefully segmented market. At the same time, while it looks like a bad idea to enter into the generic text analytics API market today, it's perhaps not such a bad idea to look at how you can build new higher-level functionalities on top of this existing NLP infrastructure: don't try to compete with what's there, but build something new using it.

Time to work on that early retirement plan.