Natural Language Engineering **24** (5): 797–803.
doi:10.1017/S135132491800027X

Industry Watch: Text analytics APIs, Part 2: The smaller players

ROBERT DALE
Language Technology Group
email: rdale@language-technology.com

Abstract

It seems like there's yet another cloud-based text analytics Application Programming Interface (API) on the market every few weeks. If you're interested in building an application using these kinds of services, how do you decide which API to go for? In the previous *Industry Watch* post, we looked at the text analytics APIs from the behemoths in the cloud software world: Amazon, Google, IBM and Microsoft. In this post, we survey sixteen APIs offered by smaller players in the market.

Text analytics APIs everywhere you look

It really does seem that a new text analytics API pops up every few weeks. In the previous *Industry Watch* post, we looked at the text analytics APIs on offer from the big players in the Software-as-a-Service marketplace: Amazon, Google, IBM and Microsoft. Those APIs are—not surprisingly, given the resources behind them—robust, well-developed and well-documented. But as we noted in that review, for one reason or another you might not want to get into bed with the big players. Plus there's always the possibility that smaller players might have fresher ideas, and be able to implement those ideas with a nimbleness and agility that escapes bigger organisations that tend to grind forward more slowly.

So, with that in mind, in this post we take a look at sixteen APIs from companies that focus on text analytics as their core business. At the time of writing, I'm aware of at least another ten text analytics APIs out there, but the total of twenty we cover in this and the previous post combined should be more than enough to give you an idea of what's on offer and what might best suit your needs. If I've missed out an API you've used and love, or—worse!—if I've missed out the API you've spent the last few sleepless months developing, drop me an email and I'll feel encouraged to cover it in a subsequent post.

The text analytics landscape

Our focus here is on cloud-based APIs offered on commercial terms via Software-as-a-Service subscription models. There are many more companies that offer software

Table 1. *The sixteen APIs surveyed here*

Company	Website
Ambiverse	`https://www.ambiverse.com`
Aylien	`https://aylien.com`
Bitext	`https://www.bitext.com`
Dandelion	`https://dandelion.eu`
Geneea	`https://www.geneea.com`
Indico	`https://indico.io`
Intellexer	`https://www.intellexer.com`
MeaningCloud	`https://www.meaningcloud.com`
Open Calais	`http://www.opencalais.com`
ParallelDots	`https://www.paralleldots.com`
Repustate	`https://www.repustate.com`
Rosette	`https://www.rosette.com`
Semantria	`https://www.lexalytics.com`
TextRazor	`https://www.textrazor.com`
TxtWerk	`http://www.txtwerk.de`
Yonder	`http://yonderlabs.com`

development services using proprietary text analytics toolsets, but they are not our concern here. We're looking specifically at companies that make available toolsets that you can use to build your own text analytics applications. We're also focussing specifically on APIs whose primary purpose is to enable the construction of applications that work on text documents, such as web pages, PDF files, mail messages or tweets; there's another class of natural language processing APIs we don't consider here that is more fundamentally concerned with building interactive chatbots. There's some crossover between these two types of applications in terms of what counts as useful functionality, and in fact a few of the vendors discussed here appear to be moving their main focus to the chatbots space; but our principal concern is the processing of documents. Table 1 lists the sixteen vendors we'll look at here.

It's important to note that every text analytics vendor provides a portfolio of functionalities, and the range of services offered will likely be a key factor in your selection of toolset. For example, if you need to do both named entity recognition and summarisation, the latter requirement already narrows down your options considerably. The most important of the capabilities offered are summarised in Table 2. This is not an exhaustive tabulation, and many of the APIs offer additional niche functionalities not listed here. It's also not necessarily the case that any two vendors who offer a specific capability mean the same thing by the terms used, and inevitably the tabulation here makes some compromises. In particular, note that the Linguistic Analysis category here is a catch-all for a wide range of functionalities: some vendors provide just part of speech tagging, a few provide some form of parsing, and a small number do open relation extraction. And just because a vendor offers a particular capability doesn't mean that they do it well.

https://doi.org/10.1017/S135132491800027X Published online by Cambridge University Press

Table 2. *Capabilities by product: ER = Entity recognition, SA = Sentiment analysis, LD = Language detection, KE = Keyword extraction, CL = Classification, SU = Summarisation, LA = Linguistic analysis*

Product	ER	SA	LD	KE	CL	SU	LA
Ambiverse	✓	×	×	×	×	×	✓
Aylien	✓	✓	✓	✓	✓	✓	×
Bitext	✓	✓	✓	✓	×	×	✓
Dandelion	✓	✓	✓	×	✓	×	×
Geneea	✓	✓	✓	×	✓	×	✓
Indico	✓	✓	✓	✓	✓	✓	✓
Intellexer	✓	✓	✓	×	✓	✓	✓
MeaningCloud	✓	✓	✓	×	✓	✓	✓
Open Calais	✓	×	×	×	✓	×	✓
ParallelDots	✓	✓	✓	✓	✓	×	✓
Repustate	✓	✓	✓	×	×	×	✓
Rosette	✓	✓	✓	✓	✓	×	✓
Semantria	✓	✓	✓	✓	✓	✓	✓
TextRazor	✓	×	×	×	✓	×	✓
TxtWerk	✓	×	×	✓	✓	×	×
Yonder	✓	✓	✓	✓	✓	✓	×

As in the previous post, in the rest of this article we'll focus on a single capability, namely entity recognition, since that's the one capability that is common to all vendors.[1] But even here, as we'll see, there's a lot of variety.

How to choose a named entity recognizer

There are many non-technical factors that will impact your choice of provider, such as quality of support and documentation, uptime guarantees, and of course price. Here, however, we'll focus on technical factors.

So let's suppose you've decided what general text analytics capabilities you want, and this has helped you filter the candidates available. Now you want to decide which named entity recognizer best fits your needs. There are a few key criteria you might want to consider.

(1) What types of entities do you want to recognize? Almost every API returns types that correspond to the basic triad of people, places and organizations, but some offer many others.
(2) Do you need to know the positions in the text where entities are mentioned? You'll want this, for example, if your application needs to show the identified

[1] If you're interested in a more detailed analysis than is possible in the space available here, covering the full range of capabilities offered by these and a number of other APIs in a more systematic fashion, see my forthcoming *Text Analytics Consumer Guide*, available at www.language-technology.com/iwtar.

https://doi.org/10.1017/S135132491800027X Published online by Cambridge University Press

entities in context, whereas if your task is just filtering documents, then knowing that something is mentioned, regardless of where, may be sufficient.

(3) Do you need identified entity mentions to be disambiguated via linking to some external knowledge source, such as Wikipedia or DBpedia?

Table 3 shows the types identified by each API,[2] and Table 4 shows how the sixteen APIs surveyed here measure up on the other two criteria. In addition to these features, some APIs provide a numerical measure of the salience of the named entity in the document; and many also deliver some numerical measure of confidence, which may be confidence that the string identified is a named entity, that the identified string is of the specified type, or that the disambiguation link provided is correct.

You'd like to see some examples, of course. There's not enough space here to show the outputs from all the APIs mentioned, so I've selected a representative sample. Given the input sentence *Jeff Bezos could be spending even more time in the nation's capital, as Washington D.C. is increasingly looking like the frontrunner to land Amazon's second headquarters*, Figure 1 shows what ParallelDots provides; this is typical of the simpler outputs. Figure 2 gives a flavour of a more complex result as provided by Ambiverse; and Figure 3 shows the results provided by TextRazor.

How well do they work?

Again, just because a vendor claims to detect entities of such-and-such a type doesn't mean they do it well. A proper evaluation is beyond the scope of this post; there are just too many variables and too many idiosynracies in requirements across use cases for a simple quantitative evaluation to make sense. But you'd certainly want to do thorough testing before making a hard-to-reverse decision about which vendor to choose. My own anecdotal testing of all the APIs listed here suggests performance varies quite significantly, so there's no shortcut around building a test set appropriate to your requirements and seeing how the APIs handle it.

Fortunately, all the vendors here offer either a free tier of usage, sometimes with quite generous call quotas, or a free trial period, so it's not hard to get a feel for how well these APIs perform. Each provides a fairly standard sign-up process via which you create an account and are provided with authorisation credentials; all the vendors provide either a RESTful API that you can use with the appropriate libraries for more or less any programming language, and/or an API wrapper that makes life easy in some particular languages.

The bottom line

So, as you'd expect, the best API for you depends on a number of factors: Is the provided documentation and support adequate for your needs? Does the API provide

[2] TxtWerk also recognizes an unusually wide range of measurements: Length, Area, Mass, Temperature, Voltage, Amperage, Resistance, Charge, Capacity, Conductance, Inductance, Magnetic Strength, Power, Energy, Force, Pressure, Frequency, Volume, Luminosity, Illuminance, Spin, Substance, Radioactivity, Currency and Time.

https://doi.org/10.1017/S135132491800027X Published online by Cambridge University Press

Table 3. *Recognized types by product*

Product	Types
Ambiverse	Person, Location, Organization, Event, Artifact, Other, Unknown
Aylien	Person, Location, Product, Organization, Keyword, URL, Phone, Email, Date, Money and Percentage; also DBpedia and schema.org ontology types
Bitext	Person, Car License Plate, Place, Phone Number, Email Address, Company, Organization, URL, IP Address, Date, Hour, Money, Address, Twitter Hashtag, Twitter User, Other Alphanumeric, Unknown
Dandelion	DBpedia ontology types
Geneea	Person, Location, Organization, URL, Email, Twitter Hashtag, Twitter User, Date and Time, Numbers
Indico	People, Places, Organizations
Intellexer	Person, Organization, Location, Title, Position, Age, Date, Duration, Nationality, Event, URL, MiscellaneousLocation, Unknown
MeaningCloud	Proprietary-type hierarchy, with Person, Location, Organization, ID, and Process as top-level categories; also Concepts, Time Expressions, Money Expressions, Quantity Expressions, Quotations and Relations
Open Calais	Anniversary, City, Company, Continent, Country, Editor, Email Address, Entertainment Award Event, Facility, Fax Number, Holiday, Industry Term, Journalist, Market Index, Medical Condition, Medical Treatment, Movie, Music Album, Music Group, Natural Feature, Operating System, Organization, Person, Pharmaceutical Drug, Phone Number, Political Event, Position, Product, Programming Language, Province or State, Published Medium, Radio Program, Radio Station, Region, Sports Event, Sports Game, Sports League, Technology, TV Show, TV Station, URL
ParallelDots	Name, Place, Group
Repustate	A total of 204 subtypes categorised under the ten top-level types: Event, Location, Person, Product, Number, Org, Science, Health, Technology and Time
Rosette	Location, Organization, Person, Product, Title, Nationality, Religion, Credit Card Number, Email, Money, Personal ID Number, Phone Number, URL, Date, Time, Distance, Latitude and Longitude
Semantria	Person, Place, Company + some subtypes
TextRazor	DBpedia and Freebase types
TxtWerk	Person, Place, Organisation, Job Title, Work, Event and Concept; also Dates and Date Ranges
Yonder	Person, Place, Organization and Misc

the functionalities you need, and at a performance level you're comfortable with? And, of course, the price needs to be acceptable—again, there's quite a remarkable variation across the usage costs of the APIs we've looked at here; see the web sites listed in Table 1 for up-to-date information.

https://doi.org/10.1017/S135132491800027X Published online by Cambridge University Press

Table 4. *NER features by product*

Product	Position	Linked data
Ambiverse	Yes	Wikidata and Wikipedia
Aylien	Yes	DBpedia
Bitext	No	None
Dandelion	Yes	Wikipedia, DBpedia
Geneea	Yes	Customer-specific, optionally based on Wikidata
Indico	Yes	None
Intellexer	No	None
MeaningCloud	Yes	Sumo, Wikipedia, YAGO and others
Open Calais	Yes	PermID
ParallelDots	No	None
Repustate	Yes	Independently sourced metadata
Rosette	Yes	Wikidata
Semantria	Yes	Wikipedia
TextRazor	Yes	CrunchBase, Freebase, Wikipedia, Wikidata
TxtWerk	Yes	Wikidata
Yonder	No	Wikipedia


```
{"entities": [
     {"name": "Amazon", "confidence_score": 0.932977, "category": "group"},
     {"name": "Jeff Bezos", "confidence_score": 0.963707, "category": "name"},
     {"name": "Washington", "confidence_score": 0.953752, "category": "place"}]
}
```


Fig. 1. Example output from ParallelDots.


```
 {"matches": [
   {"charOffset": 0, "charLength": 10, "text": "Jeff Bezos",
    "entity": {"id": "http://www.wikidata.org/entity/Q312556",
               "confidence": 1.0}},
   {"charOffset": 72, "charLength": 15, "text": "Washington D.C.",
    "entity": {"id": "http://www.wikidata.org/entity/Q61",
               "confidence": 0.9170702374574355}},
   {"charOffset": 141, "charLength": 6, "text": "Amazon",
    "entity": {"id": "http://www.wikidata.org/entity/Q3884",
               "confidence": 0.1550813585138788}}],
  "entities": [
   {"id": "http://www.wikidata.org/entity/Q61", "name": "Washington, D.C.",
    "url": "http://en.wikipedia.org/wiki/Washington%2C%20D.C.",
    "type": "LOCATION", "salience": 0.3444641035427783},
   {"id": "http://www.wikidata.org/entity/Q3884", "name": "Amazon.com",
    "url": "http://en.wikipedia.org/wiki/Amazon.com",
    "type": "ORGANIZATION", "salience": 0.2007263161218188},
   {"id": "http://www.wikidata.org/entity/Q312556", "name": "Jeff Bezos",
    "url": "http://en.wikipedia.org/wiki/Jeff%20Bezos",
    "type": "PERSON", "salience": 0.9062320200194232}]
 }
```


Fig. 2. Example output from Ambiverse.

https://doi.org/10.1017/S135132491800027X Published online by Cambridge University Press

```
{"response": {
  "entities": [
   {"id": 1, "startingPos": 0, "endingPos": 10,
    "freebaseTypes": ["/film/person_or_entity_appearing_in_film",
     "/tv/tv_actor",
     "/business/shareholder",
     "/influence/influence_node",
     "/organization/organization_founder",
     "/people/person",
     "/award/ranked_item",
     "/venture_capital/venture_investor",
     "/business/board_member"],
    "matchingTokens": [0, 1], "matchedText": "Jeff Bezos",
    "relevanceScore": 0.4242, "confidenceScore": 7.102,
    "wikidataId": "Q312556", "freebaseId": "/m/011z69",
    "wikiLink": "http://en.wikipedia.org/wiki/Jeff_Bezos",
    "entityId": "Jeff Bezos", "entityEnglishId": "Jeff Bezos",
    "type": ["Agent", "Person"]},
   {"id": 0, "startingPos": 141, "endingPos": 149,
    "matchingTokens": [25, 26], "matchedText": "Amazon's",
    "relevanceScore": 0.485, "confidenceScore": 2.896,
    "crunchbaseId": "amazon",
    "freebaseTypes": ["/organization/organization", ... 25 other types ...],
    "wikidataId": "Q3884", "freebaseId": "/m/0mgkg",
    "wikiLink": "http://en.wikipedia.org/wiki/Amazon_(company)",
    "entityId": "Amazon (company)", "entityEnglishId": "Amazon (company)",
    "type": ["Agent", "Organisation", "Company"]},
   {"id": 2, "startingPos": 72, "endingPos": 87,
    "matchingTokens": [15, 16], "matchedText": "Washington D.C.",
    "relevanceScore": 0.1533, "confidenceScore": 6.09,
    "freebaseTypes": ["/organization/organization_scope", ... 23 other types ...],
    "wikidataId": "Q61", "freebaseId": "/m/0rh6k",
    "wikiLink": "http://en.wikipedia.org/wiki/Washington,_D.C.",
    "entityId": "Washington, D.C.", "entityEnglishId": "Washington, D.C.",
    "type": ["Place", "PopulatedPlace", "Settlement", "City"]}]}
}
```


Fig. 3. Example output from TextRazor, with some details elided.

If you're faced with making a decision in this space, I hope that this post will have helped save you time in doing that. And I'd be very interested to hear of your experiences: drop me an email at `rdale@language-technology.com`.

https://doi.org/10.1017/S135132491800027X Published online by Cambridge University Press