A comparison of cloud services for natural language processing

Arnaud Cassaigne, R&D Engineer

Mondeca arnaud.cassaigne@mondeca.com

Abstract

Natural language processing technologies are becoming ever-more accessible with the emergence of software as a service. This study provides a comparison of several cloud services for natural language processing in four languages: English, French, Spanish and German. We consider four major cloud computing players: Amazon with Comprehend, Microsoft Azure with Text Analytics, Google Cloud with Natural Language and IBM Watson with Natural Language Understanding. Each offering is analyzed and rated based on context, ease of use, standard functionalities, application programming interface features, supported languages, quality of results, and price. We consider following linguistic tasks:

- language detection
- named entity recognition and concepts retrieval
- keyword detection
- morpho-syntactic tagging: part of speech, lemmatization, syntactic dependencies

Some of the offerings support the creation of customer-specific models - we did not test this feature. Our tests are based on the National Geographic article *Why Do Many Reasonable People Doubt Science?* This study shows that tested cloud services for natural language processing present similar features with slight variations. In general, using these services is easy - sufficient documentation is available, registration to the service is fast and application programming interfaces are simple. The language coverage as well as the quality level of the different functionalities is somewhat variable. In some cases, the results we obtained showed very good quality and are suitable for industrial use.

1. Introduction

This study compares cloud-based offerings for natural language processing (NLP). Our tests are based on the National Geographic article *Why Do Many Reasonable People Doubt Science?* We thank National Geographic for providing the content of the article in 4 languages: English, French, Spanish, and German.

Machine reading of human-written content is a field of artificial intelligence called natural language processing, which is at the crossroads of linguistics, semantics and automatic learning. It has now reached acceptable quality levels to leave research laboratories and be adopted across all businesses, which use it to add value to their data and increase their productivity. Frequent use cases include indexing and categorization of content, data mining, consumer feedback analysis, and information monitoring.

Using NLP technologies is becoming ever-more easy with the emergence of software as a service (SaaS) offerings. There is no need to install complex software. Performing sentiment analysis

A GIANT LEAP FOR DOUBTERS

A worker adjusts an exhibit at NASA's Kennedy Space Center in Florida. Skepticism about established science is hardly new, but the Internet has been a boon to fringe beliefs. Think the moon landings were faked? Go online—you'll find plenty of people who agree.

MAGAZINE

Why Do Many Reasonable People Doubt Science?

Figure 1. National Geographic test support article

or retrieving standardized parts of speech such as named entities or keywords does not require any expertise - you simply need to submit a piece of text to an application programming interface (API) to retrieve results. Besides, the pay-per-use pricing mechanism will limit the initial investment of an NLP project.

SaaS offerings from the 4 major cloud computing players are tested here: Amazon with Comprehend, Microsoft Azure with Text Analytics, Google Cloud with Natural Language and IBM Watson with Natural Language Understanding.

2. Methodology

Each offering is analyzed and rated based on context, ease of use, standard functionalities, API features, supported languages, quality of results, and price.

Our tests consider standard analysis of content and include the following tasks (with some variations due to differences in offerings):

- language detection of the text
- named entity recognition (NER) and concepts retrieval
- keyword detection
- morpho-syntactic tagging: tokenization, part of speech (POS) tagging, lemmatization, relationship between syntagmas

Some of the offerings support the creation of customer-specific models - we did not test this feature but we will mention it when available.

The sample text we used is limited to the first 5,000 characters of the National Geographic article *Why Do Many Reasonable People Doubt Science?* by Joel Achenbach and Richard Barnes,

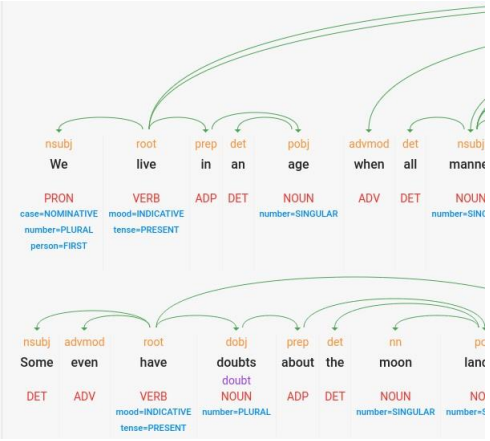

Figure 2. Part of syntactic dependency tree (Google visualization)

published in March 2015. We processed it in English, Spanish[1], French[b], and German using the available graphical interfaces. The versions published in the different languages are not word-for-word translations but still quite accurate. As the test is restricted to the first 5,000 characters of the article in each language, the English version may have more paragraphs compared to its translations (because of the word-count increase due to translation).

Performing a qualitative analysis using a limited amount of text gives a good 'idea' of what could be achieved with longer texts. Yet, to produce a scientific/definitive analysis would require testing the different offerings on well-defined text corpora and computing standard metrics such as precision, recall, and F1 scores.

It is also important to note that the sample text we used for our tests is a press article with characteristics not necessarily found in other types of content: well-constructed, syntactically correct sentences; sentences may be syntactically long and complex (subordinates, asides...); less standard characters (opening quotation marks, non-standard dashes) and sometimes more complex typographical variants (English Title Capitalization, or All Caps in the French article); is similar to the types of documents used to train linguistic and information extraction models so results are usually better than those obtained using other types of content.

The pricing simulation is based on the following simple use cases:

- Use case 1: 10,000 short documents averaging 256 bytes processed per month (this typically happens when analyzing social media content, consumer complaints, or users comments about products and services).
- Use case 2: 10,000 long documents averaging 3,000 bytes (equivalent to an A4 page) processed per month (this typically happens when analyzing larger documents such as press articles, reports or contractual documents).

For both use cases, we assumed that the documents would be analyzed in a sequence, with one service call per document, without any pre-processing operation aimed at optimizing performance; and that the same tasks would be performed i.e. language detection, NER, sentiment analysis, and keywords retrieval.

[1] The link to the Spanish article does not correspond to the beginning of the article discussed here.

[b] The article in French is not available online.

3. Amazon Comprehend

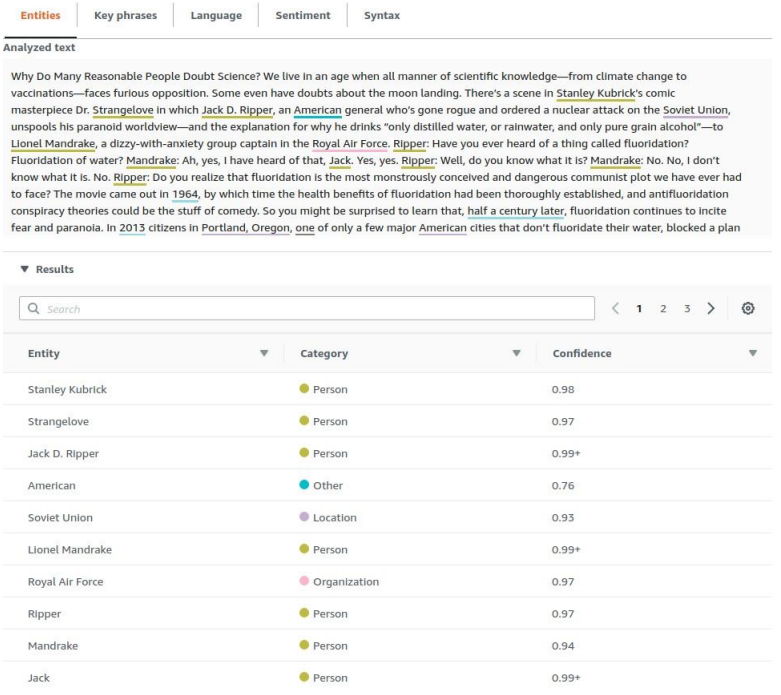

Figure 3. Amazon Comprehend graphical interface

Amazon Web Services (AWS) has embarked on natural language analysis with Amazon Comprehend, released in late 2017. It is one of the many services offered by the cloud giant. Comprehend Medical - a service dedicated to medical content analysis - was launched at the end of 2018 and is not evaluated in this test.

Comprehend supports language detection, NER, concept and keywords retrieval, sentiment and syntax analysis in English, French, Spanish, Italian, Portuguese and German. Optionally, Comprehend Custom allows the creation of specific extraction and classification models, which are not part of this test. It should be noted that the graphical user interface and the API allow asynchronous processing tasks. Besides, the API can process a document from S3 storage (storage of AWS objects).

Using this service is quite simple since you only need to create an AWS account. There is only one package coming with a free test session up to 50,000 units. One unit corresponds to 100 characters with a minimum of 3 units per call. Above 50,000 units and up to 10 million units, a unit costs $0.0001. Processing 10,000 short messages averaging 256 bytes will, therefore, cost $12. Processing 10,000 A4 pages averaging 3,000 characters will cost $120. The pricing structure is relatively simple and is well suited to a large number of short texts.

The Comprehend documentation is clear but only available in English. The graphical user interface and API limit the size of content to 5,000 characters per call, which is too low and requires larger documents to be segmented upfront. Only the Plain Text format is supported, which means the text must be pre-extracted from documents (such as XML, HTML, PDF, MSWord, etc.). The API is HTTP REST. Client libraries for this API are available in common programming languages (Python, .NET, Java, Javascript). On the privacy side, data submitted to the service can be used by AWS to improve Comprehend.

Once connected to AWS, you select the Comprehend service and choose the AWS deployment region. Not all regions are available. A test document can then be processed via the console. The graphical test interface is more than just a demo interface and certainly more usable than those of competing services. The interface highlights the hits and displays the API response. However, the interface has some bugs: it is sometimes necessary to click on the page to trigger the display of results and there is no vertical scroll bar for text display. We also received a 'size exceeded' error message while processing less than 5,000 characters.

Here are our results on the analysis of the first part (5,000 characters) of the National Geographic article, *Why Do Many Reasonable People Doubt Science?* in 4 languages: English, French, Spanish, German. The processing time for all documents is fast (1 second) regardless of the language. The language detection works correctly. Named entities are classified as 'Person', 'Place', 'Organization', 'Product', 'Date', 'Event', 'Quantity', 'Title' and 'Other'. A confidence score is provided. However, the service does not include concept retrieval. Also, there are no links to knowledge graphs (typically a Wikipedia or DBPedia URI) which is a pity. We know that 'Stanley Kubrick' is a person but we do not know who he is (a standard feature in some competing services). Results for named entities in English are very good, except for a few borderline cases. In French, the results are quite good as well. 'd'Ebola' is wrongly tagged as Organization, but with a low confidence score. Many quantifiers are detected, probably too many e.g. 'rares grandes villes' (rare large cities). 'Jack D. Ripper' is divided into 2 different entities: 'Jack D.' and 'Ripper' (this is not the case in English). Concepts are sometimes classified as 'Other' (e.g. 'fluoride', 'pandemie' (pandemic)). In Spanish, the results are very good. 'Cancer' has been tagged´ as 'Other'. In German, the results present a few errors, 2 wrong detections but with an average confidence score (0.6): 'Air-Force-General' tagged as 'Product' and 'Mandrake' tagged as 'Organisation'. The 'Ripper' occurrence was not retrieved. Too many quantifiers were detected, e.g. 'ohne Sorge' (no problem), 'alle' (all), 'viele' (many) and concepts are sometimes tagged as 'Other' entity types, for example, 'Fluorid' (fluoride), 'Ebola'.

Keyword detection, according to the Comprehend documentation, consists in finding all the nominal groups in a text and our test results show that this works perfectly well. There is a confidence index but no relevancy score that would allow to sort and filter the (many) keywords found, which is a serious limitation for further analysis of results. In English, French and Spanish, the extraction is successful even if it may cause less relevant chunking (e.g.: 'principle' in 'in principle', 'nous' (we), 'tel' (as), 'aumento' (increase)). In German, there are some problems with word and sentence tokenizations 'der Zweifler Die Skepsis gegenuber der Wissenschaft', 'doch."¨ Ripper'.

As for sentiment analysis, there is no consensus on standard metrics for sentiment scores and each offering has its logic, so you will need to study documentation to understand it properly. Nevertheless, scores associated with results sometimes seem unclear. Besides, sentiment analysis should generally be carried out on subjective texts, and may therefore not be suitable for a press article (the National Geographic's press article we used is rather factual or neutral, but dealing with "negative" subjects). Amazon Comprehend indicates a neutral feeling for English, French, Spanish. For Spanish, the feeling is neutral at 0.7 and negative at 0.2.

The syntactic analysis only provides the lexical units and their grammatical category, with a confidence index. Other competing services offer more information. In English, there are tokenization problems on probably less common Unicode punctuation characters (apostrophes, quotation marks, dashes). It may be due to a character normalization issue which should be corrected. Examples of erroneous lexical units include 'knowledge—from', or 'There's'. This incorrect tokenization leads to grammatical category errors on these lexical units. On the other hand, results for the main grammatical categories (noun, adjective, and verb) are good and useful. Other more ancillary categories are questionable (e.g.: 'why' is classified as an adverb). For French, the analyzer is misled by the all caps letter case of the first sentences: in 'AUX ETATS-UNIS, LE´ SCEPTICISME EST A SON COMBLE', 'AUX' ('in) is classified as a proper name as well as` 'EST' ('is') and 'SON' ('his'/'her'). There are some tokenization problems related to punctuation, just as in

English, e.g: ”l’evolution” (‘the evolution’) which is considered as a single lexical´ unit. But generally speaking, the part of speech tagging is good. ”Pourquoi” (‘why’) is tagged as an adverb and ”et” (‘and’) has no grammatical category. For Spanish, in the same way, the all caps noun ’INCREDULIDAD’ causes the parser to tag it as a proper name. The ”y” (‘and’) or ”que” (‘which’) do not have a category. There are some errors in the most useful categories. For example, in the sentence ’El agua fluorada causa cancer’, ’fluorada’ is classified as a verb (mean confidence index at 0.49). ’causa’ is classified as a noun with a high confidence score of 0.91 and ’cancer’ is tagged as an adjective (0.88). ‘¿Tel´ efono’ is a single lexical unit classified as a proper´ name. ”Ripper” is tagged as a verb. In German, ’Die Skepsis gegenuber der Wissenschaft’ (the¨ skepticism against science) marks ’Die’ and ’Skepsis’ as proper names. That can be due to a segmentation problem at the paragraph level since the previous sentence (the title) does not end with a full stop but is separated by 2 carriage returns. ”Haben” and ”Ah”, classified as proper names,¨ may be victims of unusual Unicode characters. Another error on ’diese’ classified as pronom in ’diese Szene’. Apart from these problems, detection in German is generally good.

4. Azure Cognitive Services Text Analytics

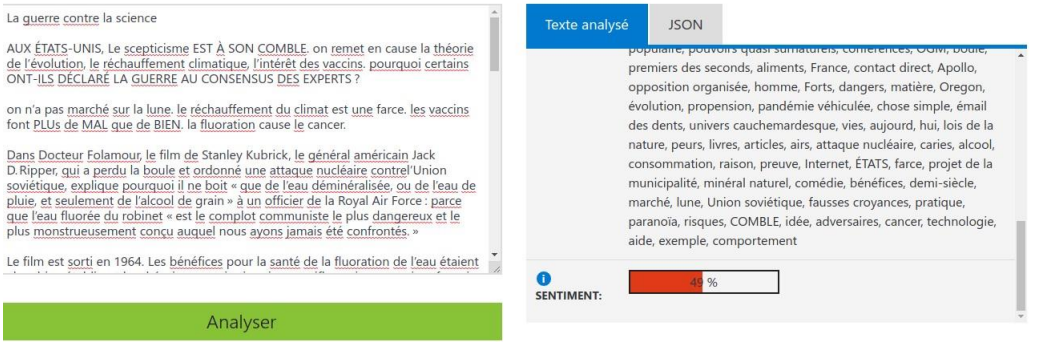

Figure 4. Microsoft Azure Text Analytics graphical interface

Cognitive Services Text Analytics is a Microsoft offering released in 2016 within its Azure cloud platform.

Text Analytics can detect up to 120 languages and supports (sometimes as a beta version only) 22 languages[2] for sentiment analysis, keywords, named entities and concepts. NER is available in English and Spanish as a stable release; sentiment analysis is available in English, French, Spanish and Portuguese. Custom language or classification models with Text Analytics are not available. However, custom automatic learning models can be developed using the Microsoft ML module or chatbot-oriented models with the Language Understanding module. Text Analytics can be used as a cloud service or launched in a Docker container. The Docker container can be deployed on any host, whether local, in the Azure cloud, or at another cloud provider. The pricing is identical for all deployment models. Moreover, the Docker container allows content to be processed locally (none of your content is copied to Azure), but the number of available functionalities is limited (language detection, sentiment analysis and keywords only).

An Azure account is required to use Text Analytics. Then you have to search for the “Text Analytics” service, which is not easy because the service is hidden in a “Plus” tab of the “Cognitive Services” directory. You must choose the deployment region. Not all regions are available.

Several packages are available, the offer is quite complex. The Free package is limited to 5,000 transactions per month. The Standard S package is $2 for 1,000 text records (of 1,000 characters

[2] European languages, Arabic, simplified Chinese, Japanese, Korean, and Turkish

each) and up to 500,000 records, then a tapering rate is proposed. S0 to S4 packages come at a fixed price per month ($75 to $5.000), and cover a volume of 25,000 up to 10 million transactions per month. Beyond that, transactions are invoiced. A transaction corresponds to an operation on a document. With the S package, processing 10,000 short messages of 256 bytes costs $75 while processing 10,000 A4 pages (of 3,000 characters) costs $226. The offer is a little more expensive than Amazon's. The main limitation of this offering is the impossibility of adapting it to specific content. It may be useful when dealing with generalist content - such as press articles - especially in English.

The Text Analytics documentation provides sufficient information but is not as clear as in other offerings. A French version, automatically translated from the English version, is available but sometimes unclear due to the poor translation. The list of languages indicates 'English' when it is actually French, which is misleading. Some resources are only available in English (API, forum, knowledge base). The graphical user interface and API impose a limit of 5,120 characters per document, which is too low and requires larger documents to be segmented upfront. It only supports UTF-8 Plain Text, which means the text must be pre-extracted from documents (such as XML, HTML, PDF, MSWord, etc.). The API is HTTP REST. Client libraries for this API are available in common programming languages (Python, .NET, Java, Javascript). On the privacy side, data submitted to the service can be used by Microsoft to improve the Text Analytics service, except when using the Docker container.

Text Analytics does not provide a graphical document analysis console as such but only a demo interface, accessible without an Azure account. This demo interface features highlighting for NER but not for keywords. The JSON output of the API is visible.

Here are our results on the analysis of the first part (5,000 characters) of the National Geographic article, *Why Do Many Reasonable People Doubt Science?* in 4 languages: English, French, Spanish, German. The processing time for all documents is fast (1 second) regardless of the language. The language detection works correctly. We tested NER with the demo interface but both the demo and JSON output do not expose entity categories. The information displayed is the Wikipedia URI and the Bing ID associated with the detected entity. There is no confidence nor relevancy score. The documentation points out that the entity categories are available with API version 2.1. This version may not be available in to the demo interface yet. Named entities are classified as Person, Location, Organization, Quantity, Date, URL and Email. NER is not available in French and German. In English, the quality of detection is not good. Named entities and concepts are misinterpreted due to a lack of verification of the context (e.g. the segments 'Do Many' are attributed to the city of Many in Louisiana, 'Ah' to Alternative History, 'No' to Norway). In another example, 'attack on the Soviet Union' is mistakenly attributed to Operation Barbarossa (an event of the Second World War). In Spanish, the results are better than in English but limited to named entities - none of the concepts were detected. Entities such as 'Moon', 'Moscow', 'Stanley Kubrick' are detected correctly and assigned to the right Wikipedia resource. However, there are still some important errors such as 'S´ı' (yes) or 'No' (no) detected as International Measurement System or Japanese Theatre No. 'Jack D. Ripper' and 'Lionel Mandrake' are not detected.

Keyword extraction aims - just like with the Amazon service - to extract nominal groups from sentences. There is no confidence nor relevancy score. The FAQ documentation points out that keywords are sorted in descending order of importance. This important information is absent from the documentation. In English, French, Spanish and German, the results are in line with the target set. Many expressions are extracted but the sorting operation allows to keep the most significant elements. In this respect, the first elements contain representative concepts of the article such as "scepticism", "scientific consensus". Sometimes, the elements extracted may be less relevant, e.g. in French "forts" in "forts de leur experience" (literally 'strong from their experience'). In´ German, an incorrect grammatical tagging may have caused the verb "glauben" to be extracted.

The sentiment analysis returns a score between 0 (negative) and 1 (positive). The English article receives a very negative score (13%) while the French, Spanish and German versions receive a score close to neutral (between 49% and 53%). Syntax analysis is not available.

5. Google Cloud Natural Language

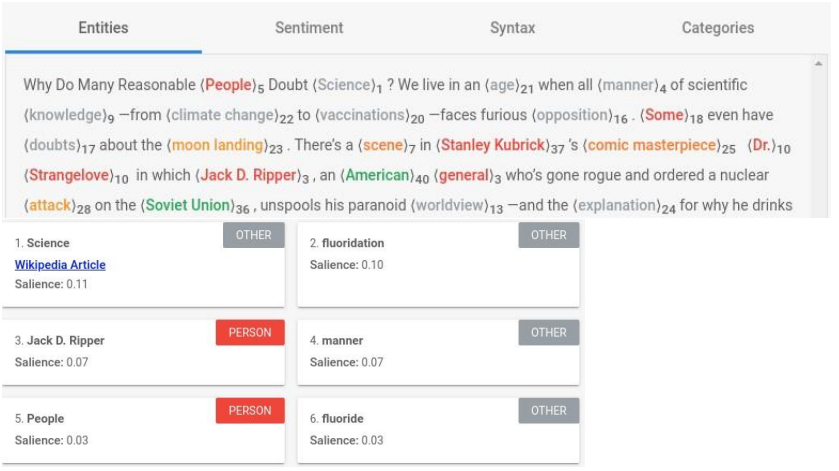

Figure 5. Google Cloud Natural Language graphical interface

Google Cloud is a comprehensive cloud computing platform. Natural Language, released in mid 2016, is the service dedicated to the analysis of textual content.

Natural Language supports content in English, French, Spanish, Italian, Portuguese, German, Russian, Simplified/Traditional Chinese, Japanese, Korean, and includes NER, concepts retrieval, syntax and sentiment analysis. There is no language detection as such but the language is automatically detected and returned in most operations. There is no keywords detection contrary to competing offers. Content classification (based on a Google taxonomy) is available for English. NER and sentiment analysis is available for English and Japanese. Specific extraction and classification models can be created with AutoML (in Beta version); we did not test this service. Batch or asynchronous processing is not available. The API can search for a document directly in a Google Cloud Storage (storage of Google Cloud objects).

A Google Cloud account is required to use Natural Language. There is only one package with a free test session for up to 5,000 units per month. 1 unit corresponds to 1000 characters for one operation. After that and up to 1 million units, a unit costs $0.001 for NER and sentiment analysis. Syntactic analysis is worth half and NER and sentiment analysis is worth twice as much. Beyond that, a tapering rate is proposed. Content classification is priced differently - it is free for up to 30,000 units per month, then $0.002 per unit up to 250,000 units. Processing 10,000 short messages of 256 bytes costs $25 while processing 10,000 A4 pages (of about 3,000 characters) costs $74.

The documentation in French is clear and pleasant to read. There are sometimes a few paragraphs left in English on a French page and some resources, such as the description of APIs, are in English only.

The API limits the size of content to 1 MB of characters per call, which is very convenient. Other limits (100,000 lexical units, 5,000 entities) are silently generating no errors but no results beyond the limit. The API supports Plain Text or HTML formats which means the text must be pre-extracted from documents (such as XML, HTML, PDF, MSWord, etc.). The API is HTTP REST. Client libraries for this API are available in common programming languages (Python, .NET, Java, Javascript). On the privacy aspect, the documentation specifies that user data cannot be used to improve the Natural Language service, but can be used for debugging or other testing purposes.

To our knowledge, there is no possible choice of the processing region. Natural Language only offers a demo interface, accessible without a Google account. This interface is quite good with NER highlighting and graphical representations adapted to each processing, e.g. the graphical

visualization of syntactic dependencies. The absence of vertical scrollbars makes the analysis of long documents tedious. Also, NER highlighting information should be manually matched by comparing a given number to its corresponding named entity category. The JSON result of the API is not accessible.

Here are our results on the analysis of the first part (5,000 characters) of the National Geographic article, *Why Do Many Reasonable People Doubt Science?* in 4 languages: English, French, Spanish, German. The processing time for all documents is fast (1 second) regardless of the language. The language detection works correctly. Natural Language detects named entities and concepts and classifies them into Person, Place, Organization, Event, Other, Unknown, Artwork, Consumer Goods, Telephone Number, Address, Date, Number, Price. Natural Language does not return a confidence score but provides a relevancy score in relation to the document itself. This is quite useful to filter out the most interesting named entities. Entities and concepts can be linked to a Wikipedia URI. In English, NER is very good. Named Entities are correctly assigned to their Wikipedia URI, when relevant. There are many detected concepts, however they are often classified as Other and rarely linked to a Wikipedia URI. Only names and not nominal groups are detected, which differentiates this type of result when compared to competing offers. The relevancy score shows pertinent results (e.g. Science) but missed other important ones (e.g. 'Belief'). There are some (rare) mistakes: 'Lionel Mandrake' is marked as Artistic Work. 'Dr. Strangelove' is identified as two different entities. In French, the relevancy score returns interesting concepts (e.g. 'Science', 'Skepticism') but sometimes less relevant ones (e.g. 'War'). Detection of named entities and concepts is not as good as in English. "marche" is a verb and should not be marked´ as a concept. 'Lune' ("Moon") is neither detected as a Place nor related to Wikipedia's Moon concept. 'farce' (joke) is wrongly detected as a Consumer good ('stuffing') as it fails to disambiguate the contextual meaning ("ridiculous situation"). 'Jack D. Ripper' is marked as 2 distinct Person entities. "Soviet Union" is detected as a Location with Wikipedia link (contrary to results in English). 'Science' (the scientific journal) is detected as an Organization, however the Wikipedia link provided points to the concept of science as a general category of knowledge. 'Fluoration' is returned as Event or Other. "organisms" in "genetically modified organisms" is classified as Organization. In Spanish, the relevancy score gives 'INCREDULIDAD' which is correct, the rest of the detection has a relevance close to 0. The detection of named entities and concepts is good but sometimes erroneous. 'RIPPER' and 'MANDRAKE' are detected in Other. "Telefono rojo" (Red Telephone) is detected in Consumer Goods. Moscu ("Moscow") is detected as a Location with´ no link to Wikipedia. "fuentes" in "fuentes de informacion" (sources of information) is detected´ in Person. "Meme" in "meme de la cultura" ("meme of popular culture") is tagged as Person. "Interstellar" is detected as Location but is well connected in Wikipedia to the movie Interstellar. "Apolo" is detected as a Person. There are some entities without type nor information (ex: 1964).

In German, relevancy captures interesting entities such as 'Zweifler' (skeptics) and 'Wissenschaft' (science). The detection of entities and concepts is quite good but there are some errors, including named entities. 'Ripper' is detected as Consumer Goods or Other. "was" is detected as Other. 'Stanley Kubricks' is classified as Other but well connected to Wikipedia. 'nein.' is detected as Other. 'einer' is tagged as a Person in 'einer der wenigen amerikanischen Großstadte' (one of the¨ few American cities). 'Frankenstein' and 'Frankenfood' are detected as Organisation. "Apollo" is returned as a Person.

Results for sentiment analysis are provided at document, sentence and entity level (entitylevel results are not exposed in the demo interface, though). Sentiments are expressed using two metrics: a score between -1 and 1, which represents a scale for negative to positive feelings, and a magnitude representing the strength of the feeling. In English, the feeling is neutral with a score of -0.1. In French, Spanish and German, sentiment analysis is not available.

The Natural Language parser provides a lot of information: grammatical categories, morphological information (lemma, gender, number, case in German) and dependencies between lexical units. In English, the syntax parser works very well. In French, except for a few errors, it is acceptable. 'AUX' is detected as singular adposition while plural. 'SON' ("his") is detected as a pronoun instead of a determiner. The analyzer may be misled by the letter case in both cases. "marche" is detected noun while verb, hence the error on the detection of entities. "Jack D.": the˙ dot is taken as the end of a sentence. In Spanish, the detection works pretty well. 'desmandado general' (a rebel General) is assigned to a name followed by an adjective while the correct tag is an adjective followed by a name. Syntax analysis in German is quite good as well. 'Ah' (ah) is¨ detected as name. "Wissen" is detected as a name instead of a verb in "Wissen Sie auch, was... ("Do you also know that..."). "nein." is detected as name while it is elsewhere detected as adverb or unknown. 'wussten' is lemmatized as 'wussen' (the correct lemma is wissen). "seltsam" is returned as an adverb but it is rather an adjective in "Ist es nicht seltsam, dass... (Isn't it strange that...)'. Additional inaccuracies are found in 'den USA, in Portland, Oregon, einer der wenigen amerikanischen Großstadte ohne fluoridiertes Trinkwasser, wurden im Jahr 2013 entsprechende¨ Plane der Stadtverwaltung abgeschmettert' ('In the United States, in Portland, Oregon, one of¨ the few major American cities without fluoridated drinking water, the municipality's project was rejected in 2013.'): 'einer' is detected as masculine nominative whereas it is a feminine dative related to 'Stadt'. The sentence structure is slightly incorrect. 'ihrem' is detected as pronoun while determiner (ihrem Wasser / their water).

Natural Language features a categorization capability based on a specific classification scheme. This feature works for English texts, but not for French, Spanish or German. Natural Language classified the National Geographics article in the 'Health' category with a 0.52 relevancy score (it failed to detect science or belief as the main subjects).

6. IBM Watson Natural Language Understanding

The company AlchemyAPI was founded in 2009 and developed a SaaS offering for Natural Language Processing. in 2015, IBM acquired AlchemyAPI, integrated it into the IBM Watson cloud offering and changed its name to 'IBM Watson Natural Language Understanding' later in 2016.

Natural Language Understanding handles content in Arabic, Simplified Chinese, Dutch, English, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish and Swedish with very different levels of functionality such as NER and concept detection and relationships, keyword retrieval, sentiment and emotion analysis, syntax, semantic roles (following a subjectaction-object pattern), and content categorization using a specific classification scheme. Natural Language Understanding can also automatically extract metadata from HTML files. The detection of relationships between entities is an interesting feature, yet we could not test it as it is unavailable from the demo interface. IBM Watson Knowledge Studio can be used to create specific extraction and classification models (we did not test this service).

An IBM Cloud account provides access to the Natural Language Understanding service. Simply choose the deployment region and the package. The 6 proposed centres are located in North America, Europe and Asia-Pacific regions. The SaaS service proposes a free package limited to 30,000 NLU units per month, a standard package at $0.003 per NLU unit (up to 250,000 NLU units per month). A tapering rate is proposed for over 250,000 NLU units. An NLU unit

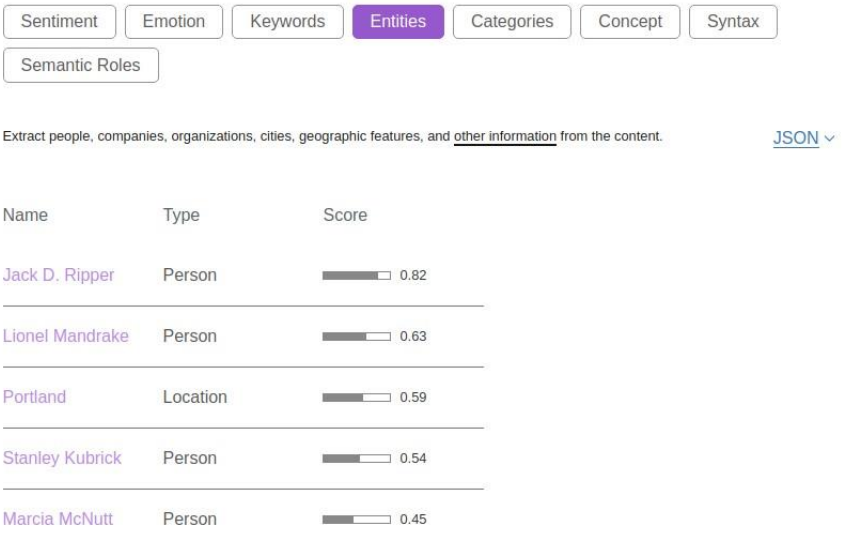

Figure 6. IBM Watson Natural Language Understanding graphical interface

is a block of 10,000 characters per API function. It is also possible to contact IBM for a pricing adapted to your data security requirements. With the standard package, processing 10,000 short messages of 256 bytes costs $118 and processing 10,000 A4 pages (of about 3,000 characters) also costs $118. This is the most interesting offering for long, or even very long documents. But adapting it to a specific business case might add to the cost and require a real technical-financial study.

The documentation of Natural Language Understanding is clear and in French but the API documentation is only available in English. The API limits the size of content to 50,000 characters per call, which is quite reasonable. The API supports Plain Text or HTML formats which means the text must be pre-extracted from documents (such as XML, HTML, PDF, MSWord, etc.). The API is HTTP REST. Client libraries for this API are available in common programming languages (Python,.NET, Java, Javascript). On the privacy side, the API logs requests and results by default to improve the service, but you may optionally add the 'X-Watson-Learning-Opt-Out' header in the request to prevent it.

The Natural Language Understanding GUI is a functional demo interface which does not require an IBM account. The JSON result of the API is available. However, no highlighting is available. It is therefore necessary to look up in the JSON result for named entities linking to DBpedia. There is a bug for semantic roles - only the first sentence is displayed graphically. However, the JSON output is complete.

Here are our results on the analysis of the first part (5,000 characters) of the National Geographic article, Why Do Many Reasonable People Doubt Science? in 4 languages: English, French, Spanish, German. The processing time for all documents is fast (2 seconds) regardless of the language. The language detection works correctly. The named entities categorization depends on the version. NER classes (Person, Location, Organization, Date, Quantity,...) are divided into categories and subcategories resulting in approximately 450 categories! Categorization is therefore a lot more granular than with competing offerings. We analyzed named entities and concepts, which are 2 separate functions for Natural Language Understanding. Concepts do not have a specific category. Entities and concepts are provided with a relevancy score and a link to DBpedia when available. There is no confidence score. In English, named entities and concepts retrieval is very good with correct DBPedia URIs. We found some errors: 'rewards—but' attributed to Location (perhaps a problem of character normalization), 'Frankenfood' attributed to Location, 'Apollo moon' attributed to a Geographical Entity. In French, concept retrieval works pretty well even if the classification list is quite

short (8 concepts). Concepts come with their relevant DBpedia URIs (e. g. 'Science', 'Belief'). NER is also quite good but only one entity has a link to DBPedia ('France'). There is a bug in the JSON output - entity labels are often truncated (e.g. 'Docteur Folamou', 'gen´ era', 'am´ ericai'). Entity types are correctly detected, except for a few´ errors: 'Apollo' is marked as Person, ':' as "IP Address", 'Ebola' as Organization, 'GMO' as Organization. In Spanish, NER works well but not as good as in English. Links to DBpedia are available. Some unexpected characters are shown in the JSON output display (probably just a display bug). Other errors were found: 'Nnnn-no' marked as Person, 'engano' ('trickery') marked˜ as "Health condition". Most concepts are correctly retrieved with their DBpedia URI, with a few errors - e.g. Estados Unidos (United States), Stanley Kubrick should have been detected as named entities, not concepts. The list of proposed concepts is short. In German, named entities contain many errors and only the 'NASA' entity has a DBpedia URI. Internet links to the German DBPedia did not work at the time of the test. Examples of errors include: 'Stanley' Location, 'Zweifler' ('skeptical') Location, 'Komplott' ('plot') Person, 'Paranoia' Person, 'Gene' Person, 'einem' ('one') Measure, 'Forschungsergebnissen' ('search results') Location. 'Jack D. Ripper' is separated into 2 distinct Persons. Detected Concepts are interesting, but there are only just a few.

In English, French and Spanish, keyword extraction provides a selection of correct nominal groups, sorted by decreasing relevance. There are many results but they can be filtered by score to keep only the most interesting ones. In German, results are mainly single words and not nominal groups (e.g. 'Jahr' ('year'), 'Offizier' ('officer').

Natural Language Understanding returns a sentiment score between -1 (negative) and 1 (positive) at document and sentence level. It can also provide emotion analysis in English. In English and German, content is considered negative with scores ranging from -0.54 to -0.71 respectively. Spanish is rather negative with a score of -0.37 and French is neutral (0.0).

The Natural Language Understanding parser returns lexical units, their grammatical category and lemma. Natural Language Understanding also provides semantic roles made out of subjectverb-object triples. This corresponds to one of the syntactic-dependency-type functionalities. In English, we noted that the syntax analyzer is misled by the capitalization of the title "Why Do Many Reasonable People Doubt Science?" - all words, except for 'Why', are categorized as proper name. We also spotted other errors: 'when' is an adverb, 'faces' in 'faces opposition' is a name, 'gone' is an adjective and 'rogue' is a name in 'who's gone rogue', 'unspools' in 'unspools his paranoid worldview' is a name. For semantic roles, the results are quite good but we noted a weak sentence segmentation and misinterpretation in 'Strangelove in which Jack D. Ripper, an American general who's gone rogue and ordered a nuclear attack on the Soviet Union, unspools his paranoid worldview (...)': the subject is 'a nuclear attack', the verb 'order' and the object 'on the Soviet Union'. Another example: 'Mandrake: Ah, yes, I have heard of that, Jack', the subject is 'I', the verb 'have' and the object 'heard of that'. In French, syntax analysis and semantic roles are not available. In Spanish, Natural Language Understanding provides semantic roles but no syntactic analysis. Note that the JSON response of the API includes strange additional characters in the result sentences, probably due to a display problem in the interface. Some results are clearly wrong. For example, "La era de la INCREDULIDAD", the subject is "la", the verb "era" and the object "de la incredulidad". Another example, '¿Por que motivo personas razonables ponen´ en duda la razon", the Verb is 'qu´ e'. In German, Natural Language Understanding provides the´ semantic roles but no syntactic analysis. Again, there are some errors. In 'Verschworung Theorien¨ verbreiten sich immer schneller', the subject is 'Verschworung Theorien', the verb is 'verbreiten'¨ and the object is 'sich'. In this case the sentence contains a pronominal verb. In the sentence 'Mandrake: ¨h, ja, doch, ich hab davon gehort, Jack.', there is no subject, the verb is ' geh¨ ort',¨ the object is 'Jack'. Again in 'Sollten wir uns Sorgen machen', the subject is 'Sorgen', the verb 'machen' and the object 'uns'.

Natural Language Understanding can provide content categorization based on a proprietary classification scheme. In English, the content is classified into / health and fitness, / health and fitness / dental care, / health and fitness / disease. In French, the content is classified in / health and fitness

/ dental care, / law, govt and politics / politics / health and fitness / disease. In Spanish the content is classified in / health and fitness / dental care, / law, govt and politics / politics / food and drink. In German the content is classified under / health and fitness / dental care, / food and drink. Just like with the Google Cloud Natural Language categorization, the National Geographic article falls under the 'Health' category, whereas the main topic of the article is science and belief.

7. Conclusion

The cloud-based offerings for natural language processing we tested present similar features (with slight variations). In general, using these services is easy - sufficient documentation is available, registration to the service is fast and APIs are simple. The language coverage as well as the quality level of the different functionalities is somewhat variable. In some cases, the results we obtained showed very good quality and are suitable for industrial use.

Table 1. is a summary of available functionalities and languages, per offer.

Table 1. Available functionalities and languages per offer

	English	French	Spanish	German
Named entities and concepts	A, G, I	A, G, I	A, G, I	A, G
Keywords	M, I	M, I	M, I	M
Sentiment Analysis	A, M, G, I	A, M, I	A, M, I	A, M, I
Simple syntax	A, G, I	A, G	A, G	A, G
Advanced syntax	G, I	G	G	G

A: Amazon Comprehend M: Microsoft Azure Cognitive Services Text Analytics G: Google Cloud Natural Language I: IBM Watson Language Understanding
Simple syntax: lexical unit, grammatical category, lemma Advanced syntax: syntactic dependencies or semantic roles

Table 2. is a recapitulation of prices for our two use cases.

Table 2. Pricing per use case

	Amazon		Google	IBM	Azure
Monthly cost for 10,000 short documents averaging 256 bytes	$12	$75	$25	$118	
Monthly cost for 10,000 long documents averaging 3,000 bytes	$120	$226	$74	$118	

To map your planned use case to a price, you will need to factor in the key elements of an auto-tagging project: document size, number and frequency of calls and the possibility of grouping/splitting content. Variations in these factors will substantially affect the price. Pricewise, Amazon and Google are attractive for short and concise documents, whereas IBM has been designed for larger documents.

We recommend to assess your project needs in terms of features and languages before choosing between these natural language processing services. Your project may also require additional capabilities such as server virtualization, image categorization, automatic learning, etc. that can be offered or not on the same cloud platform. Only some offerings propose custom linguistic features and models that may be needed for advanced users. Another important element is the rapid pace of

improvements of these technologies and new offerings releases will lead to better results in the near future - a small-scale test right before you get started will confirm your options.

References

Achenbach J. & Barnes R. Why do many reasonable people doubt science? *National Geographics* in English, Spanish, French and German

Amazon Web Services. Comprehend cloud service

Microsoft Azure. Text Analytics cloud service

Google Cloud. Natural Language cloud service

IBM Watson. Natural Language Understanding cloud service