Machine Learning & Deep Linguistic Analysis in Text Analytics

Text analysis is commonly done using two distinct types of approaches. Machine learning approaches, which are based on statistical methods, are the most common. Linguistic approaches, which are based on knowledge of language and its structure, are far less frequently used.

These two types are frequently seen as alternative, competing approaches, particularly in Big Data analysis. This view represents a major obstacle to the progress of the Big Data industry. The two approaches should not be seen as competing but as complementary and cooperative approaches that, when properly combined, result in the most effective way of extracting high-quality insights from big data. We strongly support this view at Bitext.

The relevance for Big Data

Big data is a booming business. The analysis of big data is proving to be an effective tool to support the business decision making process. By exploiting the structure of data, capturing the relationships between data elements (objects, concepts and actions), Big Data extracts key insights such as customer preferences, attitudes, product features, selling points, sales leads, legal strategies, employee satisfaction?, etc.

Text is a significant part of Big Data. The nature of big data is extremely varied. It can be roughly divided into numerical and textual data. Looking at the contents of big data, there is a wide variety of data types: numbers (spreadsheets mostly), database records (clients, providers), machine-generated texts (logs), human-generated texts (social media, blogs, news…) A significant percent of these is text generated by humans. Examples of this type of human-generated text are e-mails, social media, news articles, online reviews, etc. There is an enormous amount of valuable information in this data, but traditional approaches to Big Data analysis based on classifying data (categories, positive/negative opinions, etc.) are only able to extract superficial information because they don't understand the structure of language.

Still machine learning is the leading approach. Given the fundamental difference between numerical and textual information, it might be surprising to find that machine learning techniques, which are based on mathematical and statistical frameworks, have become the most commonly used tools for text analytics and mining.

Machine learning and Linguistics are not competing approaches. In this context, the view that machine learning and linguistic approaches to text processing are competing approaches has become widespread. However, this is a misconception largely brought about by lumping together machine learning approaches for big data analysis, which can work on top of either type of approach, and machine learning or statistical approaches to dealing with text in general, which are indeed incompatible with linguistic approaches. This misconception is not based on solid reasoning, machine learning and linguistic approaches can work together: linguistic approaches are ideal for understanding language and providing it with the structure that machine learning needs to extract accurate insights from text.

Why is Machine Learning used for Text Analytics?

Statistical approaches are a quick but limited option. Machine learning and, in particular, statistical approaches to text processing have probably become trending for two reasons. First and foremost, they are mathematical tools that are well understood by computer scientists, which makes them the natural and comfortable choice when dealing with a problem; by turning a problem into a classification, clustering or modelling task, engineers without deep knowledge of linguistics can obtain immediate results. The relative success of statistical approaches to speech recognition and machine translation only helped to cement this view, even as linguistic techniques matured and became more efficient.

Talent trained in Linguistic approaches is scarce. Secondly, linguistic approaches require extensive knowledge of both the science that studies language (linguistics) and computer science. This discipline is known as Computational Linguistics and, unfortunately, people trained in this area are not frequently found.

What are Machine Learning's drawbacks?

Machine learning ignores sentence structure. Most commercial text analytics solutions based on machine learning do not consider the structure of sentences, opting instead for a "bag of words" approach which cannot capture the relationships between words.

Ignoring the structure of a sentence can lead to various types of errors: false positives such as mistaking "durable good" as expression of positive sentiment, or incorrectly detecting similarity in phrases such as "social security in the media" and "security of social media". Not taking into account the effects of certain types of words leads to the inability to deal with common language phenomena such as negation (where "I do not like this phone" is clearly a negative opinion, even if it does not contain an explicit negative word such as "dislike") or conditionality (where "I would recommend this phone if the screen was better" is not a positive statement even though it contains the word "recommend"). It also makes it impossible to deal with granularity in language, where a sentence like "the screen is wonderful but I hate the on-screen keyboard" contains two distinct opinions that must be evaluated separately.

Machine learning requires training datasets. Training of machine learning systems is another major issue. Correctly selecting the appropriate training set is not trivial; in statistical sentiment analysis, for example, overfitting of the training data can often lead to incorrectly labelling neutral concepts (such as "Harvard" or "Stanford") as being positive or negative.

Big Data, by definition, implies variety; and variety in text data means different types of input text, ranging from formal language (news or reports, for example) to informal language (emails, call center transcripts, customer surveys, social media data). This variety poses a major challenge to the training paradigm of machine learning, since every type of data has different training needs. Besides, Big Data is processed for different business purposes: improve client service, prevent churn, generate sales leads, prevent default of loan payments… Again, every business purpose has different training needs.

Training needs are a limitation of machine learning. These two factors put together, variety of input text types and variety of business purposes, pose a challlenge for the principles in which machine learning is based. By definition, machine learning solves problems it has been trained for. Coping with the variety of text types and purposes poses a major challenge for this ad-hoc, training-based approach, in terms of the need for creating training data, a process that by definition is performed by hand and is costly and error-prone. As a result, the need for training is a major obstacle for the success of machine learning as a way of extracting insights from text.

Finally, the result of training a machine learning system is a "black box": if the training set causes the system to incorrectly classify a sentence or detect the wrong sentiment, there is no easy way to tweak the system to fix the error other than additional training or complete retraining. In either case, this involves gathering new manually annotated data which can be expensive.

How does Deep Linguistic Analysis deal with these issues?

Linguistics understands the structure of sentences. Deep Linguistic Analysis (DLA) employs knowledge about language (grammars, ontologies and dictionaries), which allows it to deal with the structure of language at all levels (morphology, syntax and semantics).

By taking into account the structure of language, DLA can deal with complex phenomena like negation and conditionality accurately, especially in complex cases where "I don't plan to buy this product" and "if I don't buy this product today I can buy it tomorrow" have a similar wording but entirely different meanings.

Correctly understanding the structure of a sentence also allows DLA to provide granularity. For example, in the sentence "the screen is wonderful but I hate the on-screen keyboard", we can detect the presence of multiple opinions without combining a positive and a negative opinion into a single neutral one. More importantly, it allows us to identify the topics or concepts being discussed, which means we can identify that the positive opinion is about "screen", and the negative opinion is about "on-screen keyboard".

DLA's is mature enough technology for business use, outside of research laboratories. The computational grammars, ontologies and dictionaries efficiently describe general language structure and content, and as a result can be applied to different kinds of text, from short Twitter sentences to long legal documents.

Additionally, DLA employs a "glass box" approach, where rules are encoded explicitly, and improvements can be made easily by adding new rules or modifying existing ones, all with predictable results.

Linguistics can solve the unsolved part of the problem. Typically, a DLA engine can be configured to analyze a wide variety of text, based on the commonalities that all human language expressions share. For example, it is feasible today to define grammars and lexicons capable of analyzing different types of text news or social media sources. Besides, DLA is naturally suited to be customized for tackling different business applications: prevent churn, generate sales leads, etc.

The Good News: Linguistics and Machine Learning are compatible

In summary, DLA provide two major advantages:

- DLA properly solves the problem of structuring different types of text and for different purposes
- DLA converts unstructured text into a structured output that machine learning can use to extract insights

And, as a result, machine learning and DLA can complement each other to deliver the promise of accurate and reliable data discovery that the market is expecting now.

DLA is designed to provide structure to unstructured text. By splitting the architecture of Big Data text analysis into two phases, introducing a linguistic structure extraction step that generates a rich and accurate representation, allowing machine learning to extract insights from actual features, which is the task that it naturally excels at.

Case Studies

The following Case Studies are good instances of how Bitext DLA helps accomplish specific business objectives, like identifying which aspects of my services are valued by users and which ones are not, or identifying sales leads for the banking sector. These objectives are achieved just by analyzing the content of data sources commonly available to any company and aligning this analysis with specific business goals.

Identifying weaknesses and strengths in Enterprise Feedback Management. A provider of enterprise feedback management to hotel chains monitors web reviews for these hotels. Customer feedback is gathered from all public sources like TripAdvisor or Expedia, to collect millions of reviews per day in different languages. Traditional text analysis systems can only classify opinions as positive or negative. Bitext DLA platform can provide the hotel chain with higher granularity like WHAT aspects of the experience (room, staff, customer service, etc.) are perceived positively or negatively and WHY (rooms are too small, the staff is pretty rude, etc.). Besides this WHAT and WHY are categorized according the business priorities, like customer care or brand perception so the bulk of opinions are easy to read and act upon: what specific aspects of the customer experience need improvement, or which positive aspects can be actively used in advertising.

Generation of sales leads in CRM. A German commercial bank is interested in discovering new business opportunities, specifically for their loans business line. The bank wants to monitor any source of public news, looking for stories that hint for need for financing such as plans for product launches, expansion of production facilities, new investment lines or mergers and acquisitions. Traditional text analysis systems are capable of classifying news into different categories or themes, but they cannot extract granular information that is critical for the system to be effective: which company is launching a new product, or which company is being acquired and by whom; when and where are these events happening. DLA extracts this detailed event knowledge (actor, action, object) that is the base to detect sales leads, letting the bank contact prospective customers before their competitors do and anticipating their own existing customer's needs.