01 - INTRODUCTION

Artificial Intelligence (AI) is hot right now – in the press, at conferences, on TV. Along with self-driving cars, virtual reality, and 3d printing, people can’t get enough of AI. This is due in part because it is associated with cutting-edge technology, new applications that think like humans, and people possibly being made obsolete and losing their jobs.

The term AI is linked with specific technologies and techniques such as Machine Learning, Deep Learning, Natural Language Processing, Latent Semantic Indexing, and others. The truth is many people don’t fully understand what these terms mean, how they are applied to software solutions, or why they should care.

At Seal, we use many of these technologies in a unique way to power our discovery and analytics software. It is good for users of Seal to have a basic understanding of what they mean, and the value they bring to our platform.

02 - WHAT IS ARTIFICIAL INTELLIGENCE?

Much of the world was introduced to AI with IBM Watson, the computer that in 2011 beat Jeopardy champions Ken Jennings and Brad Rutter on TV. But, is this form of AI the technology that will power all applications going forward? Can it really learn and make cognitive decisions for us? Will it be smarter than humans and take our jobs?

The term AI is an umbrella term for many technologies that work together to provide solutions that can be taught through example. In Europe, the preferred term is Machine Intelligence, but in all cases, the term has been applied so broadly, many AI researchers feel it has become vague and carries little value.

03 - COGNITIVE AI vs. NARROW AI

It is helpful to further define the term AI by looking at “cognitive AI” and “narrow AI.” Cognitive means the ability for a software/hardware system to process and understand information, and make decisions across many topics. Cognitive AI takes sight, sound, motion, knowledge, etc., and processes this information to make a determination or create output of some kind. This emulates the thinking humans do, and is the source of much of the AI hype. Cognitive AI sparks a vision of robots doing the work humans do.

Narrow AI is the application of similar AI technologies to specific use cases, where humans work to teach a system to evaluate information, and identify or categorize specific data such as specific types of unstructured data in a contract document in the case of Seal. Narrow AI has two forms, a focused capability, that after teaching, can perform a specific task, and broader question and answer systems like IBM’s Watson, Salesforce’s Einstein and others. This second type attempts to provide answers to human generated questions in natural language, rather than the classification of discrete items. These Q/A systems are targeted at a different problem than a solution like Seal, and need different types of teaching.
According to researchers, cognitive AI for practical applications is a long way off. Narrow AI, on the other hand, is being effectively used right now in areas such as medical diagnosis or legal case analysis. AI is built on technologies that learn by example, so the work to train systems is limited to “narrow” functions now vs. the incredibly diverse training needed for cognitive AI. Practical AI is a set of algorithms within a Machine Learning framework focused on a narrow, and often well defined domain that performs specific tasks involving data inputs and outputs. Several articles claim feasible uses of cognitive-based AI are out as far as 2050 for the “singularity moment,” but many people follow the hype and are being misled about what is real and what is fiction.

When it comes to taking jobs, the AI “robots” may take jobs, but it won’t be based on replicating the reasoning ability of humans to accomplish complex tasks. It will, and is, being seen now in areas such as data processing, extraction of information from unstructured content (what Seal does), and repetitive decision making. These areas tend to be lower value work for humans, but provide real benefit to organizations when they are automated with narrow AI.

04 – WHAT IS MACHINE LEARNING?

Machine Learning is the set of technologies that makes systems seem “intelligent” and capable of learning through example. ML is the true enabler and fastest growing aspect of AI. It is pervasive, and we all touch ML (and therefore AI) in obvious ways now, such as when you ask Siri for directions to the nearest restaurant, or Facebook suggests a friend for you to tag in an image. When writing, spell check suggests spellings of words, and when texting your IM app uses ML techniques to predict and provide possible word choices as you text along.

The learning capability in software is the ability for it to take examples of something, including language, pictures, data, etc. and store a numerical representation of them for comparison against new examples in the future. When more diverse examples are evaluated by the system, it gets better and better at identifying similar examples from new data. One way it can learn is called “supervised” learning, where humans teach it by providing examples, and then telling the system if it is doing well or poorly with its results, a bit like teaching a child. But the real excitement about the future of ML is with “unsupervised” learning, where the system is provided only limited input by humans, and builds parameters – essentially learning – by processing data and combinations of data on its own.

So how does it do this? It is really a math function, where the inputs and outputs are known but the parameters are not. So the whole problem is to build a model of this mathematical function in some automatic way that finds the best parameters to achieve desired results. It uses algorithms, or combinations of algorithms (which are structured processes and sets of rules) to process and compare its examples with new data. There are many types of algorithms, including neural networks (modeled after how the brain works), decision trees, Bayesian belief networks, K-Nearest Neighbors, self-organizing maps, case-based reasoning, instance-based learning, hidden Markov models, and other various regression techniques.

The trick to successful ML is applying these various algorithms/models, in a way that optimizes the results for specific tasks and objectives, and researchers and data scientists are always working to develop new and clever ways to determine the optimal models for a particular application.
05 – NATURAL LANGUAGE PROCESSING: PUTTING THE LANGUAGE IN AI

While AI/ML is used to build learning systems to process information, applying that capability to unstructured data, specifically the type in verbal and written human language requires another set of capabilities. NLP is simply the part of AI that involves language, and by combining ML, computational linguistics, and computer science, NLP allows a machine to understand language, a task that in the past had been exclusively reserved for humans.

NLP is being applied to more areas in recent years, including sentiment analysis – or the ability to identify subjective information in texts – text classification, automatic categorization of text (such as certain contract terms in the case of Seal), information extraction, and conversational agents.

In truth, NLP can’t understand text but simulates understanding. To do so, the system must either use the grammatical rules of a natural language or create an algorithm that will infer these from data (text) – or even use both for different tasks. There are two main approaches in NLP, with the first being symbolic, relying on the symbols (i.e. the words, word pattern and some metadata about these) to determine if a word is a noun, verb, or adjective, but also whether in a given context it is a subject, object, or predicate to decipher language. This approach relies on a lot of hand-crafted lexicons and word-lists to specify what the words mean. In English, a rule may look like “if a noun precedes a verb, then this noun is the subject.”

An alternative method is the statistical approach, in which the words are tracked by numbers, counts, and frequencies. Based on this numeric data, we can use many methods from statistics, mathematics, and probability theory to infer a rule such as: “in English it is most probable that an adjective precedes a noun, but vice versa in French.” These rules, induced or derived with the help ML or Deep Learning need training data in the form of labelled text with the desired output.

In recent years, hand-crafted rules that work on words have widely been abandoned in favor of the statistical approaches as they seem to generalize much better over previously unseen data. Rules are cumbersome to maintain for very complex tasks while the semantics, or meanings of words or phrases can be encoded as numbers. The rest is to have the ML or DL engine provide the output in a given context.

06 – LATENT SEMANTIC INDEXING: FINDING THE MEANING OF TEXT

The trouble with language is that humans use vast combinations of words to transfer meaning. Words themselves can mean very little in some cases, with combinations of words being the only way to understand what is being said or written. Humans can decipher meaning from various combinations of words naturally and effectively, a task much harder for a computer system designed to consider texts literally. LSI is the technique that helps systems understand the gist and meaning from various combinations of words, even across various locations in a document. Seal uses advanced LSI for the identification and tracking of meaning in contracts, even if the words being evaluated are not part of the same sentence or paragraph.

LSI is actually a form of unsupervised ML using example documents. It uses mathematical models to identify patterns in the relationships of words contained in an unstructured collection of text, and uses categorization to classify types of documents and their concepts.
LSI is valuable for returning search results in contracts that aren’t exactly the phrases used in the search by identifying important words/phrases that occur in conjunction with each other in text. For example, if the terms “performance” and “warranty” are closely related in numerous documents, when a user searches “performance,” LSI will track the association and return documents containing the word “warranty.” With contracts, terms and provisions can be written in many ways, and LSI helps Seal find and extract the right data.

07 – HOW ACCURATE IS ML?

With the increases in practical use of narrow ML to solve business problems, it is necessary to understand the acceptability and usability of outcomes. At Seal, the application is the identification of contract data, and there is a clear way to measure and manipulate the accuracy and acceptability of results to ensure the technology is meeting business objectives.

At Seal, when a customer purchases the software, it is not a process of loading disks, hitting a “magic button,” and out comes all the contract data needed for any use case. As Seal is a teachable system based on ML, humans must provide examples and then it goes out and finds those things it “knows.” Seal has been extensively trained and extracts large amounts of contract data immediately, but if a customer needs to understand specific or unique components in their contracts, the system will need to be given examples and then start the process of learning on its own. But no ML system is perfect, and errors in processing can be made.

The two primary measures of accuracy in narrow ML-based applications such as Seal are “precision” and “recall.” These look at how well an ML system is providing useable results and are represented as percentages. These metrics can be tuned by the training and inspection work that goes into an implementation.

Precision measures how well a system retrieves relevant documents in a return set. It is measured by dividing the number of documents accurately returned by the extraction by the total number of documents returned in the extraction. For example, say we want to find all contracts in 10,000 that contain 30-day payment terms. We start with a sample of 100, and using automated search we get a result of 10 contracts with 30-day terms. But after human review, we find that of the 10 returned, only 8 really have 30-day payment terms, and 2 have payment terms shorter or longer than 30 days. The precision of this return set is .8 or 80%. If all 10 contracts in our return set were proven to have payment terms of 30 days, that would result in a precision score of 1.0.

Recall measures the percentage of documents in the return set which meet the search criteria, against all documents in the portfolio which meet the search criteria. This measure focuses on which documents in the portfolio were missed in the search. Continuing with our example of searching for contracts with a 30-day payment terms, you will remember that after running our automated review we achieved 80% precision. Now, after manual review of all of the 100 documents in our sample set, we find 12 documents of the 100 actually had the 30-day payment terms we were looking for. Therefore, our automated review returned 8 of the actual 12 across the portfolio having the right terms, which results in a recall score of .66.

In this example, the user would determine if these precision and recall scores are adequate for the objectives of knowing 30-day payment terms. If so, they would run the system against their entire portfolio of 10,000 contracts, and expect similar results of .8 precision and .66 recall. If those numbers are not adequate for the purpose, then a bit more tuning and training can be used to bring them up before the run against the entire population.
As mentioned, the system is very “tunable” to balance and adjust the measures of precision and recall. To achieve high precision, a user wants to include only those items that exactly fit the criteria to keep the percentage of correct items in the return set very high. For high recall, a user will want to include items that have some variance from examples to ensure few items are missed. The Seal platform is weighted towards recall, meaning not missing any documents with the needed criteria. When needed recall and precision scores are both high, Seal will soon be providing a sliding scale in the UI to capture the best balance of weighting for the needed outcomes.

08 – THE POWER OF SEAL

One of the most important aspects of Seal’s implementation of narrow AI is the combination of ML, NLP and LSI technologies to create a high performing platform for contract data extraction. A second notable capability is the flexibility with various approaches to conduct an extraction within one platform. To be more specific, some jobs may be simple and require basic keyword searching to be resolved very quickly. Others might need deeper analytics with either heavy or light training. This flexibility results in greater efficiencies and value due to the reduction in time, effort, and cost for data extraction.

Seal’s Contract Analytics (ScA) is a pre-trained and tuned system, and is considered “reactive.” To run a search with ScA, users only need a single example, and due to the work Seal has done with the platform, it can achieve strong precision and recall within a very short time – from seconds to minutes – with the one example. Reactive searching is the first place to go when an urgent request comes in, and it can detect contract information using ranges of values, extracting discrete entries such as dates, parties, places, numbers, percentages, etc.

Seal offers a second approach which is called a proactive method, or one which requires additional examples to be provided by users. This was introduced with Seal’s User Driven Machine Learning (UDML) so business users can interface with Seal directly, providing positive and negative examples before a run is performed. UDML is slower to learn, but has the capability to “generalize” the data, meaning to summarizing data points in order to make better predictions over new and unseen documents. UDML also is good for provisions stretching over page breaks, headers and footers, or consisting of short sentences.

UDML requires a minimum of 50 examples going up to approximately 200 to effectively train the system. This training consists of giving it both positive and negative examples of sentence-level clauses, and letting it learn by statistical analysis of the examples. Proactive approaches, using UDML have the potential to provide very high degrees of recall and precision while generalizing over new data.

09 – THE PROGRESSION OF AI

The technology of Seal, and other AI solutions has been advancing rapidly over the past 10 years or so. Not as much with the math aspects involved, but more so with hardware and ability to process the high volumes of computations in an acceptable amount of time.
There is an increased use of graphics processing units (GPUs) being used in place of traditional CPUs to dramatically boost computational power and accelerate Deep Learning methods in analytics and engineering applications. To add to this, development and advancement of various types of neural networks, those models which power ML solutions, have given developers more options and processing efficiency, resulting in a much faster turnaround in learning and results. While 10 years ago, it might have taken weeks to get results back within a ML training process, that kind of timeframe is no longer even remotely acceptable.

In the next 10 years, we will see further convergence in hardware and algorithms. Chips will be produced with ML and DL algorithms right on the silicon, or perhaps within the quantum computing space. Alongside GPUs, we are starting to see IBM, Google, and Facebook all releasing hardware targeted at Deep Learning. This will continue to significantly impact the future of information processing, accuracy, and confidence, and should make AI solutions even more important and viable across many more business functions and processes.

10 – SOURCES

- [http://www.wired.co.uk/article/machine-learning-ai-explained](http://www.wired.co.uk/article/machine-learning-ai-explained)
- [http://www.cse.msu.edu/~cse960/Papers/LSI/LSI.pdf](http://www.cse.msu.edu/~cse960/Papers/LSI/LSI.pdf)

To learn more, visit [www.seal-software.com](http://www.seal-software.com)

CONTACT US:

[www.seal-software.com](http://www.seal-software.com)

**Americas HQ**
1990 N. California Blvd. Suite 500
Walnut Creek, CA 94596. USA.
T: +1 650 938-SEAL (7325)

**European HQ**
1-2 Hatfields, Waterloo,
London
SE1 9PG, United Kingdom.
T: +44 203 735 9898