expert.ai

Essential Techniques and Criteria for Successful Hybrid NL Use Cases

Introduction

Language powers business. It fuels processes, shapes internal and external communications, and offers insight into the markets that surround us. We spend enormous amounts of time immersed in the language of our work, whether we're processing and interpreting documents, searching for information or engaging with customers and each other.

AI-based natural language processing and understanding (NLP/NLU) technologies enable us to comprehend language and extract data from documents, manage interactions in natural language (e.g., chatbots) and process unstructured information at speed and scale. **As the volume of language continues to grow exponentially, NLP/NLU technologies provide a key competitive advantage for enterprises in every industry.**

AI-based NLU technologies leverage different techniques that can be divided into two segments: symbolic and machine learning/deep learning. The latter is commonly mistaken to be the only way to automatically process text.



SYMBOLIC AI

Symbolic artificial intelligence is a method of artificial intelligence research based on high-level, "human-readable" representations of problems, logic and knowledge.



MACHINE LEARNING

Pure ML, also known as "black box" AI, uses mathematical algorithms to identify statistical patterns and extract information. It does this based on inference rather than actual knowledge.

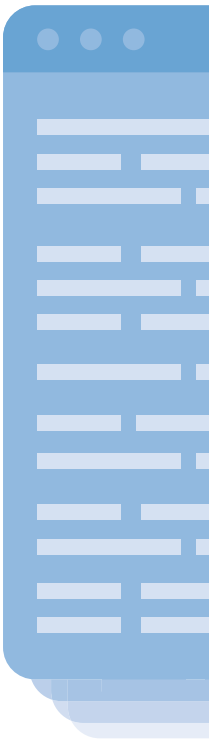
Though symbolic is lesser known than machine learning as an AI technique, it has proven essential to successful natural language processing and natural language understanding solutions. There are three specific use cases that come to mind.

The Value of Symbolic AI in NLP/NLU-Based Applications

Classification

Automatic classification of text according to a predefined taxonomy, or hierarchy of topics, is a common and efficient method of organizing information for retrieval. Symbolic and ML-based techniques differ in the way the automatic classification is reached:

- **Symbolic AI-Based:** This approach enables you to create business rules that capture the elements of a document and assign it to a category. These rules can be written manually or generated automatically (with the possibility of manual validation).

  The availability of a rich knowledge graph significantly streamlines the process of writing these rules by making them explainable (to know why a document is assigned to a category), flexible and powerful for effective classification. This technique also offers an important advantage in that performance is not dependent on the data available because humans guide the steering wheel in the so-called training phase. This allows you to:

  - create rules even when limited data is available up front through generalization ensured by human intelligence and knowledge.
  - leverage the knowledge graph, with its concepts and links to other concepts, to enable more efficient development of rules and easier refinement.

  As a result, you can reach and maintain the level of performance required and at the same time, for example, fully understand why a set of documents was or was not classified correctly (explainability).

- **Machine Learning-Based:** Classification systems based on machine learning require a set of documents to be trained on. The larger the set of documents representing each node of the taxonomy, the better. This training set must be pre-annotated by humans so that the algorithm can learn what is in the document and identify examples of how specific content is paired with specific tags or taxonomy nodes.

  For example, if we want an ML-based system to recognize documents related to finance, it must ingest a large quantity of finance-related documents that have already been annotated as "finance," together with a similar number of documents tagged as "non-finance." After processing many documents by the different features identified within them, the system will, at some point, recognize when a document belongs to the finance category.

The level of accuracy a trained system provides varies based on the number of documents used during the training phase and the coverage of specific jargon in those documents. The system must also be retrained frequently to maintain the same level of quality.

Extraction

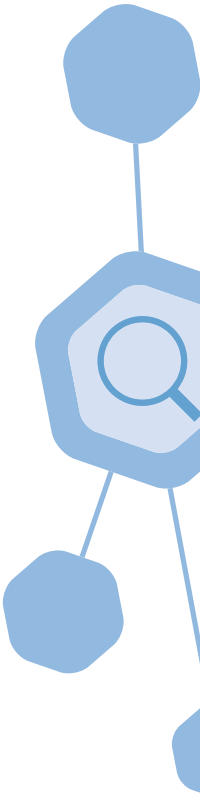
Extraction refers to the process of obtaining specific data from text, normalizing it into a structured format and archiving it in a database or other repository. The knowledge graph embedded in symbolic technologies offers significant advantages which include:



- **Virtually Unlimited Extraction Capabilities:** The automatic identification of concepts and meaning during text analysis, as well as the availability of additional information on each concept within the knowledge graph make it possible to extract data beyond the standard named entities offered by other technologies without any additional work. For example, lion, cat and tiger are identified in the knowledge graph as both "animal" and "feline." Therefore, you can extract felines mentioned in a document for further processing directly from standard text analysis.
- **Shorter Time and Lower Costs:** Expert.ai offers shorter implementation times and, therefore, lower implementation costs by automatically performing concept identification and extraction.
- **Higher Precision and Recall in Extraction:** Thanks to out-of-the-box recognition of tens of thousands of concepts enabled by its rich knowledge graph, expert.ai uniquely collects the additional contextual information necessary to increase the accuracy of extracted data.
- **Triples Identification and Export:** The entities and concepts extracted by the system can be linked through many kinds of relations automatically recognized during analysis then exported in RDF or other standard formats. For example, the system can automatically identify all relations between entities then further classify them based on the kind of relation (e.g., communication, movement, etc.).
- **Relevant and Rich Tagging:** The deep analysis provided by expert.ai offers different criteria by which you can rank elements of text, such as semantic relevance and role played by each concept in the text. This ensures more flexibility and increases the value of the automatic tagging of the analyzed content.
- **Easy Implementation of Custom Ontologies:** The richness and completeness of the core knowledge graph as well as compliance with standards (such as SKOS) makes it effortless to use custom ontologies for extraction when compared with other techniques such as ML.

Search

Search refers to process of locating specific information within a text, often from a large corpus of documents. There are a number of techniques that can be used for search, but those that can understand the meaning of each word in the indexing phase and leverage a knowledge graph with all the embedded relationships between concepts offer more unique and sophisticated features such as:



- **Conceptual Search:** A rich knowledge graph not only enables retrieval of documents related to a specific query, but of documents that include conceptually similar information. For example, an expert.ai search for "car" would retrieve documents that include the keyword "car", as well as those within the "car" concept including automobile, SUV, sedan, convertible, BMW, Mustang, and more — all without any manual customization of the index or the query. In addition, a conceptual search for the word "jaguar" would retrieve only documents related to the car, rather than those related to the animal.
- **Natural Language Search:** A symbolic AI approach enables human-like interpretation of intended queries by accounting for every different way to express them including slang, abbreviations and even misspellings. On the other hand, a traditional search application would process the individual words and ignore the implied question.

  By understanding the intended question, you return only the set of documents that explicitly answer it. This is especially valuable when seeking to maximize the performance of something like a chatbot or FAQ search.
- **Rich, Customized Metadata:** The richness of the results derived symbolic text analysis simplifies the implementation of advanced search capabilities such as faceted search. It also enables other unique features such as rich document summarization.

  The rich analysis also enables the implementation of customized ranking mechanisms by using flexible and parametric criteria. For example, R&D could have the results of a news search on its competitor ranked by its relation to product innovation while users in finance could have the same items ranked by closeness to critical business events, M&A activities or any number of other topics.

12 Criteria to Evaluate Before Selecting an NLP/NLU Technique

There are several criteria you should use to thoroughly evaluate the ideal NLP/NLU technique for your specific use case. For both symbolic and ML approaches offer varying levels of each. These should all be properly accounted for in your buying decision.



Accuracy: To evaluate which technique best addresses your specific use case, first consider how accurately the technique can perform it. Both symbolic AI and the latest generation of ML models, especially large end-to-end networks, can be very accurate under specific conditions.

To reach high accuracy, ML models require an abundance of supervised data (a rarity for a typical enterprise use case). Then, to maintain it, they need the language in the documents processed during the production phase to be very similar to that processed during training.

Symbolic AI-based models can always reach the level of accuracy required through continued refinement of the low-code rules on which they are based. However, the decreasing return of these incremental improvements cause a trade-off that each organization must consider to ensure the project delivers optimal ROI.



Scalability: Scalability depends on the effort required to annotate data or set up knowledge through domain experts. A symbolic approach, such as expert.ai, is often perceived to be less scalable as it requires language experts to develop low-code rules. This perception ignores at least three important facts that, when combined, should be objectively considered during scalability evaluations:

- Creating and training a supervised set of documents requires significant time and manual effort directly from subject matter experts (SMEs).
- SMEs are more expensive, critical and rare in an organization than the people who can develop the low-code rules.
- ML systems cannot be set and forgotten about. They must be retrained every time the performance degrades.



Latency in Production: Many NLP/NLU-based applications must operate in close to real time. This is true for chatbots and any other AI-augmented customer care applications, but it is also relevant for applications such as document processing automation. Thus, it is important to note that ML-based systems, especially those reliant on large networks (Large Language Models), can be extremely slow and fall short of the speed requirements of certain use cases.



Computational Cost or "Carbon Footprint": In an era where enterprises are required to be good citizens and prioritize sustainability, the carbon footprint of an AI investment is a paramount consideration. Large Language Models, and even general ML models, can be very expensive both in terms of training (and retraining) as well as the costs of CPU/GPU, memory and time. Thus, their carbon footprint is typically orders of magnitude higher than that of symbolic techniques.



Explainability / Interpretability: Being a good corporate citizen also implies that you avoid bias and ensure equal treatment of all people. A system that performs tasks and reaches results that are not auditable presents a level of operational risk that should not be acceptable by the enterprise.

ML systems have the core characteristic of being black boxes, meaning they lack both explainability and interpretability. Even the new approaches to deep network explainability — promoted as solutions to the black box issue — simply follow a reverse-engineering pattern without granting real control and are not scalable in real-world scenarios.

Symbolic techniques, however, are white boxes that leverage rules and, in the case of expert.ai, a rich knowledge graph — all elements that are fully auditable and understandable by humans.



Robustness to Data Scarcity: ML-based systems, and the related Large Language Models, require large data sets for quality fine-tuning that must be highly relevant to the specific use case. **Most NLP-centered uses cases in the enterprise do not have large enough training sets of documents to ensure these systems can reach the expected level of accuracy.**

This problem doesn't exist for symbolic AI systems. These systems only require a sample as small as a few dozen documents for a specific use case and the natural human capability of SMEs. This capability allows SMEs to generalize the observations derived from the analysis of these small data sets and develop the system so it can reach the highest level of performance for the specific use case.



Robustness to Noisy Data: Noisy data not only refers to corrupted or distorted data, but also to the very common scenario of a low signal-to-noise ratio — a measure of how relevant collected data is to the described phenomenon. **Subtracting noise from data is a critical step in implementing NLP/NLU-based solutions, particularly when regularizing and generalizing the terms within a text.**

Using explicit, symbolic domain knowledge, such as what expert.ai provides with its knowledge graph, can be very helpful in performing this task — one very complicated for pure ML systems given their lack of a knowledge-based language understanding.



Use Case Applicability: Use cases that organizations struggle to collect sizable enough training sets for or those that require out-of-the-box meaning understanding and relationship extraction functionalities (e.g., legal contract or policy review) can be addressed only by explicit symbolic representations and deep linguistic analysis. This makes ML-based systems practically unusable in these situations.



Accountability Full-Control "Human in The Loop": Only explainable-by-design and interpretable-by-design symbolic models offer humans full control during the development and so-called training phase. In being based on an open, interpretable set of low-code rules, **symbolic systems create a simpler way to correct for insufficient performance through adaptation and modification of the sets of rules by a subject matter expert.**

Black box ML models only offer the opportunity to add more data to the training set without an opportunity to interpret why the results have not reached the level of performance required.



Robustness to "Real-World Variations" of Data Distribution: A common scenario for any NLP/NLU-based use case is one in which many characteristics of your data change over time. This creates a gap between the training data and the data an NLP/NLU-based system faces in production. This common scenario forces the system to adapt to the changed conditions.

For ML-based systems, this simply means retraining from scratch. However, retraining requires availability of another big and relevant training set, time, energy consumption, and more. In this scenario, symbolic systems prove superior due to their explicit, symbolic domain knowledge and robustness derived from regularization and generalization. In addition, **symbolic systems maintain an advantage by being explainable and requiring less effort to adjust to data distribution variations.**



Document Fidelity: In the document world, fidelity refers to how similar the format of the document required by the application is to the format of an actual document. High document fidelity is critical to the speed and simplicity of implementing an NLP/NLU solution because it enables annotation directly on an original document (e.g., pictures in a document) and provides relevant contextual information to the annotator.

An explainable symbolic approach allows your system to "interact" with the input by adding a prediction (classification or extraction) and an explanation following the format of the original document — all without affecting its integrity or context.

On the contrary not all ML-based systems (e.g., transformers) can maintain fidelity throughout every step of the NLP/NLU workflow. Therefore, they must be correctly evaluated to ensure that their inherent limitations do not create insurmountable issues.



Deployability: In general, ML-based and symbolic systems present similar scenarios for this final criterion. However, large language networks and services offered only via cloud only make it possible to integrate with an application via external APIs rather than be deployed on the organization's network or private cloud. For data privacy needs or other use case specific needs, an on-premise installation is sometimes the only feasible option. This makes the flexibility of deployment of expert.ai another important aspect to consider.

Why is Expert.ai Unique?

Natural Language Understanding

Expert.ai technology is composed of several integrated elements that serve to analyze texts and process natural language. Its main components are:



Knowledge Graph: The knowledge held by expert.ai technology is organized within a knowledge graph. This is the core element of expert.ai technology, and it is available in 12 languages. It enables AI systems to assign specific meanings to each word it analyzes within a text through a process called disambiguation, similar to how humans apply "common knowledge" to their reading of a text.



In the same way human knowledge improves when it learns new things, expert.ai's knowledge graph is enriched and improved by a mechanism of supervised/unsupervised learning with current and domain-specific information (see below for more details),

Contextual Understanding: During document analysis, expert.ai uses a retention technique to determine the semantic context of a text, enable disambiguation and extract meaning with the best possible precision.

Thanks to the aforementioned elements, expert.ai can distinguish the correct meaning of words and expressions in context and automatically associate attributes of more general terms that are conceptually linked to these words. For example, it automatically comprehends that the terms "SUV" and "sedan" share similar attributes derived from being a kind of "car." It also identifies the relationships between concepts in a text.

Both features derive from its ability to perform different levels of linguistic analysis (morphological, grammatical and syntax analysis) in conjunction with semantic analysis/word disambiguation.

This human-like ability to read text and understand language is different from any other linguistic engine or model because it increases accuracy, speed and the ability to handle complex text.

Content Representation



The initial output of expert.ai's analysis is a conceptual map of the text, where:

- Each concept expressed in the text is uniquely identified, regardless of the words used to represent it.
- Each agent, or subject, is associated with the action carried out.
- Each object is connected to the related action.

Thanks to this representation, the content of a document (e.g., main topics, most relevant sentences, dates, etc.) is given structure, which enables it to be stored in a repository and used for tasks such as indexing, classification, summarization and extraction.

Conclusion

The increasing quantity of information that must be processed and understood on a daily basis, as well as the need to augment the capacity of the subject matter experts working on complex and critical processes are just a few organizational challenges that NLP/NLU technologies can help to address.

Symbolic NLP/NLU technologies offer the advantage of a deep, flexible and contextual understanding of information when compared to ML-based technologies. In addition, they comply better to other requirements, from explainability and sustainability to robustness of data and data distribution variation, positioning it as the best option for real-world implementations.

Finally, symbolic is not only a high-performing technique for the typical activities of information retrieval, classification and extraction, but also a promising approach for developing more advanced applications (both enterprise and consumer) and integrating with the most advanced predictive models.

Companies working on these use cases should always evaluate which technique better fit their needs. The expert.ai platform allows to contrast compare and blend the techniques for a best possible outcome.

Get Started

Interested in learning how NLU AI will transform your company? Get started here.

See what expert.ai can do for you!

expert.ai

About us

Expert.ai is the premier artificial intelligence platform for language understanding that augments business operations, scales data science capabilities, simplifies AI adoption and provides the insight required to improve decision making throughout organizations. The expert.ai brand is owned by Expert System (EXAI:BIT), that has cemented itself at the forefront of AI-based natural language solutions across Insurance, Banking, Publishing, Defence & Intelligence, Life Science & Pharma, Oil Gas & Energy, and more.

www.expert.ai **Contact Us**

©expert.ai 2021