LYMBA

The Power to Answer™

Introduction to K-Extractor™

K-Extractor™, short for Knowledge Extractor, is Lymba's core NLP system. All things Lymba go through the pipeline, and you will be amazed by the knowledge it produces.

Why a pipeline and what does it do?

Natural Language Processing is a way to break down text into data a machine can understand. Lymba does this by analyzing the text document in sequential steps of increasing complexity.

The K-Extractor™ Pipeline

Document Preprocessing

Analyzing structure of documents. Multiple types supported, including image only files like PDFs with tables. Text normalization: spell-checking, auto-correction, special characters, social language.

Text Segmentation

Determining the boundaries of words and sentences.

Part-of-speech Tagging

Labeling words as nouns, verbs, adjectives, etc.

Concept Extraction

Identifying entities and concepts through:

- *named entity recognition - identifying entities like proper nouns*
- *collocation identification - a phrase is one unit, like "hot dog" or "water under the bridge"*
- *event detection - like birthdays mergers, and acquisitions*
- *temporal expression - inferring dates and times like yesterday*

Word Sense Disambig.

Determining which sense of the word is used, like bass. The fish, a sound, or the instrument?

Syntactic Parsing

Breaking down each sentence into its grammatical components to establish structure, like direct objects, verbs, pronouns, adjectives, etc.

Semantic Parsing

Identifying 26 standard, out-of-the-box semantic relationship types with semantic calculus to establish custom semantic relationships.

Coreference Resolution

Contextualizing nouns and their relation to other nouns in the surrounding text.

Text Classification

Classification through machine-learning processes:

- *sentiment classification - identifying and labeling sentiment*
- *provisions tagging - tagging sentences and identifying provisions*
- *topic detection - identifying text as associated with a topic*

RDF/TriX Representation

Converting knowledge data into a graph database format.

LYMBA

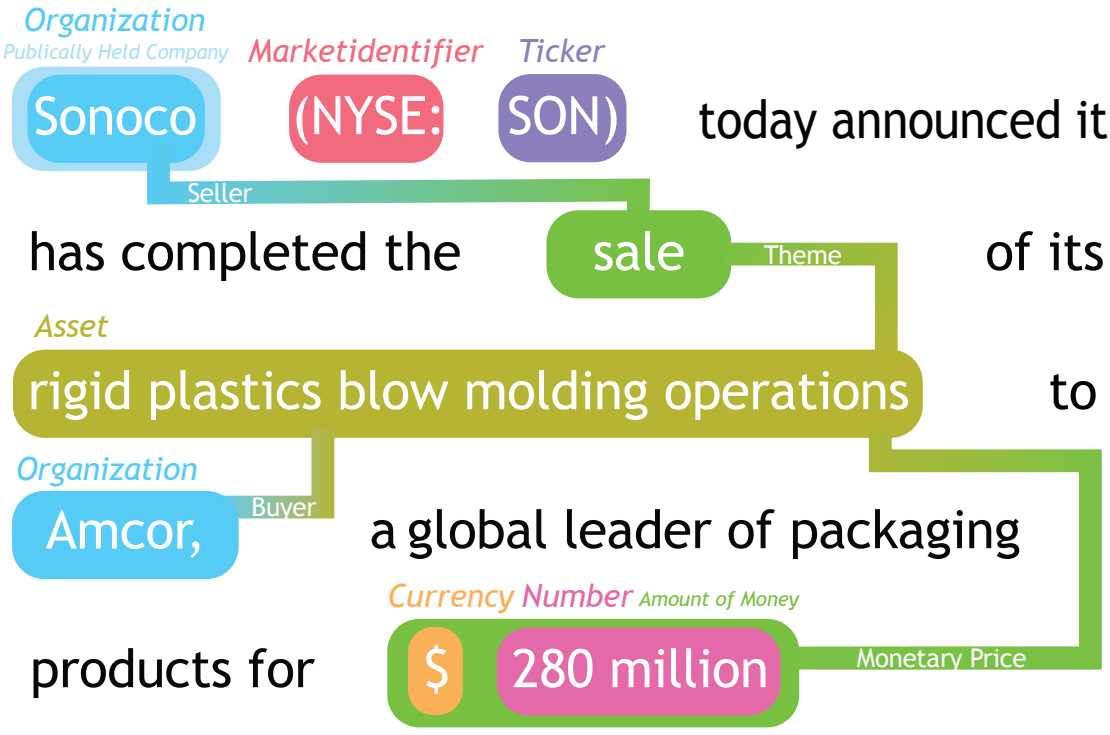
Lymba's Semantic Calculus is how we extract knowledge with high accuracy using the NLP pipeline. It is fast and easy to customize. After a text is tagged with 26 out-of-the-box semantic relationship types, the Semantic Calculus enables domain-specific relationships to be inferred by leveraging the training data. This combination makes the Lymba NLP pipeline unique and provides unparalleled levels of accuracy across any domain.

Named Entity Extraction

Entities are extracted from the documents, using:

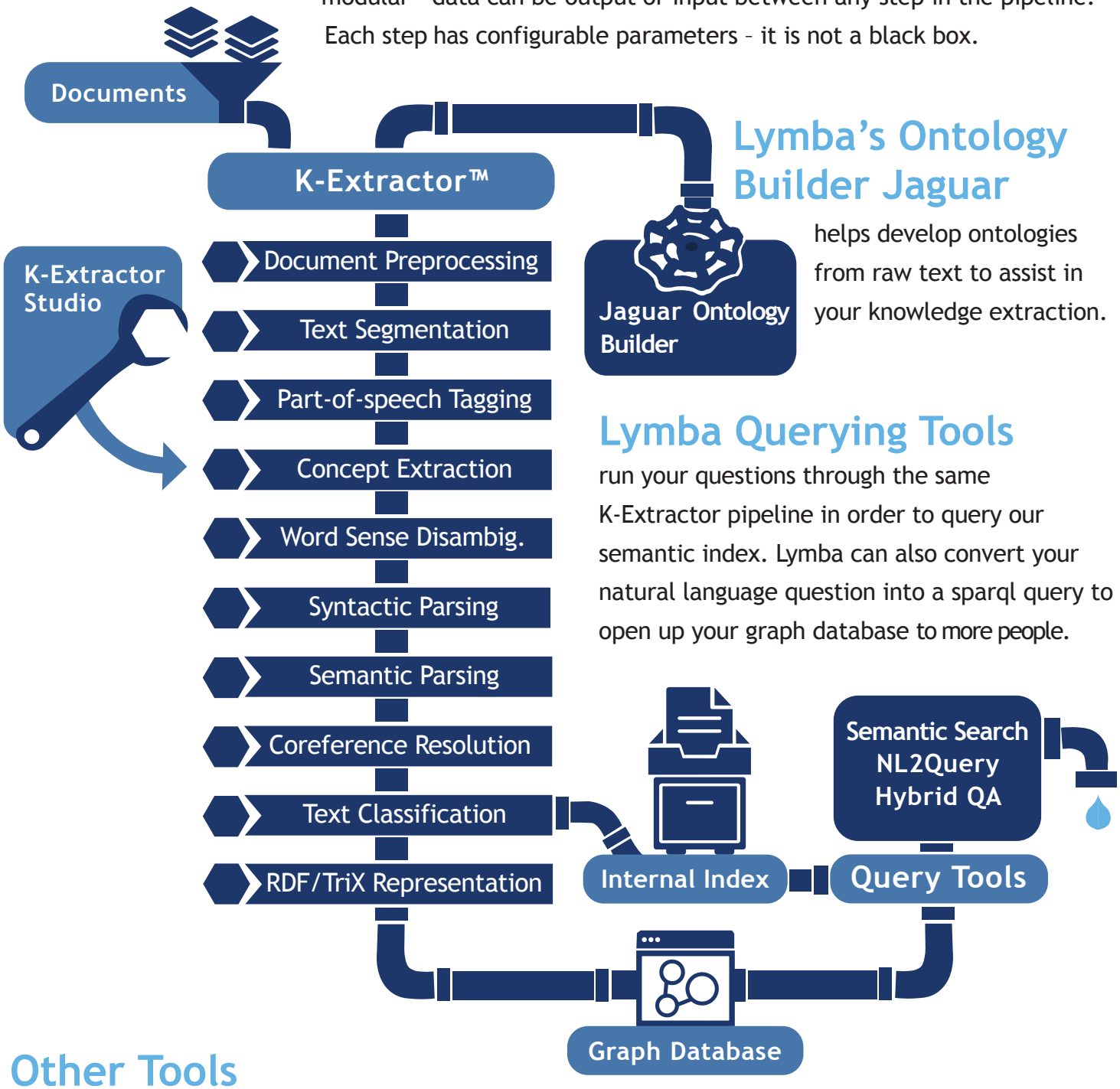
- Ontologies with tools like Jaguar
- Lexicons
- Fuzziness controls (Paris, Texas or Paris, France?)
- Extraction of dates, names, locations
- Post-processing for overall consistency
- Lymba-only tools for lexicon and rule-learning

Knowledge Extraction Example



LYMBA

K-Extractor Studio gives you the power to make changes in the pipeline. Each step is modular - data can be output or input between any step in the pipeline. Each step has configurable parameters – it is not a black box.



Lymba's Ontology Builder Jaguar

helps develop ontologies from raw text to assist in your knowledge extraction.

Lymba Querying Tools

run your questions through the same K-Extractor pipeline in order to query our semantic index. Lymba can also convert your natural language question into a sparql query to open up your graph database to more people.

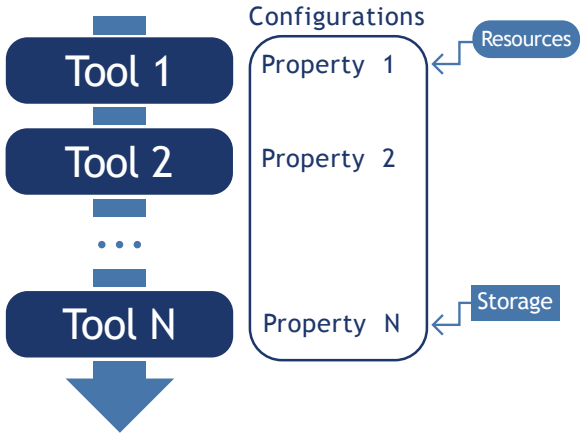
Other Tools

Graph store integrations, machine learning & classification, word embeddings, folksonomy and metadata generation, literature recommendations and document summarization, foreign language phrase mapping, text-to-ontology building, reasoning capabilities, timeline creation.

LYMBA

Pipeline Tool Configurations

Information is configured using tools from the pipeline. These tools use models/rule paths, third-party end points, a number of iterations, and lazy/eager execution. They can also identify what to include/exclude from the output as well as use strategy/machine learning methods.



Configuration Process

Lymba differentiates itself in the NLP space with it's powerful K-Extractor pipeline and also a unique configuration process. In order to extract semantic knowledge across any domain, we configure the system with rules generated from ontologies and source documents. We automate this step in two phases: 1) with a tool to rapidly build out an ontology and 2) with a tool to automate the configuration of the pipeline rules with that ontology.