Parabole

Why context is key in knowledge extraction

MD. ISHAQ
ROHIT KHANDEKAR

Contents

Introduction

Natural language processing, or NLP, is a branch of artificial intelligence that has deep implications on how computers and humans interact. Today, enterprises from all domains are seeking to leverage NLP to unlock the wealth of information stored in unstructured data, to help automate business processes and to ultimately gain a competitive advantage. Some of the central problems in NLP include machine translation, knowledge extraction, summarization, and question answering. To solve such problems, we not only need to understand the nuances of human languages, but also need to advance state-of-the-art computer algorithms and technology to be able to recognize and absorb pertinent information from vast amounts of data. In this article, we focus on the role of context in knowledge extraction.

Importance of context in knowledge extraction

Automated extraction of key concepts, entities, data elements and their inter-relationships from a document is one of the central problems in knowledge extraction and has numerous business applications. To be able to extract such knowledge, accurate with respect to the author's intended meaning, one needs a good understanding of the context in which a given document is written. According to the Merriam-Webster dictionary, context is "*the parts of a discourse that surround a word or passage and can throw light on its meaning*." Without the backdrop of a well-defined context, a word, a concept, a sentence or even a full document may be misinterpreted. Consider, for example, the two sentences below where the word "bank" has distinct meanings depending on the context.

> "I enjoy working in a bank."
>
> "I enjoy working near a river bank."

The first sentence refers to a job or a profession in a banking institution, while the second emphasizes the performance of some activity near a river bank.

Identifying the proper context, however, is often challenging because of varying levels of abstractions used in expositions. The context may change from document to document in a corpus, from paragraph to paragraph in a document, from sentence to sentence in a paragraph, or even from phrase to phrase in a sentence.

Parabole presents the full context in a hierarchical manner – with domain-level, document-level, paragraph-level and sentence-level contexts.

Hierarchical contexts in domain-specific learning and document-level analysis

We represent domain-level context via "knowledge graphs" which encode the vocabulary of domain-specific concepts, their definitions, hierarchies, properties and multi-faceted relationships and a collection of rules by which more relationships can be inferred. We extract such domain-level context via natural language processing of domain-specific corpora of documents that are carefully chosen. We refer to this phase of context extraction as *domain-specific learning.*

During the *document-level analysis* phase, we use the backdrop of domain-level context to analyze a given document, on the fly, by doing further natural language processing. We set the document-level and other lower levels of contexts by extracting key concepts, named entities and data elements, their meanings, hierarchies and inter-relationships from the document and vet them against the knowledge graph. We may also choose to "incremental learn" our knowledge graph by augmenting it with new knowledge extracted from the document.

Many times, the contexts may be fuzzy and we must allow for multiple interpretations. We handle such ambiguity by using a probabilistic model of entity classification during the analysis phase.

A probabilistic model of entity classification

Consider, for example, classifying a named entity "Summit Bank" mentioned in a document. Without additional context, such a bank can be either:

a) A financial institution [in finance], or

b) A snow bank, river bank [in geology], or

c) A blood bank, gene bank, sperm bank [in biology and medicine].

In order to disambiguate, we take the document-level, paragraph-level or sentence-level contexts into account. Even when there is insufficient information regarding an entity, Parabole's probabilistic model, based on a rich knowledge-base of concepts, entities and their categories, infers entity-types accurately. Similar to named-entity recognition and classification, a probabilistic model can be used at the document and paragraph levels to identify context distributions.

Extracting semantic contexts at sentence level

Linguistic analysis of vocabulary terms may not be enough for NLP algorithms to correctly apply learned knowledge. To successfully apply learning, the algorithms must "understand" and "use" the semantics of every vocabulary term within the context of a given document. For example, consider two sentences:

> "Under US GAAP, gains and losses from AFS assets are included in net income."
>
> "Under IFRS, gains and losses from AFS assets are included in comprehensive income."

Both sentences refer to 'gains' and 'losses' from some form of income, but the resultant information or knowledge needed to be learnt is entirely different between the two due to differing semantics. It is this combination, encompassing both probabilistic and

deterministic methods within Parabole's cognitive analytics solutions, allows our algorithms to truly understand the meanings within a selected text.

Conclusion

In this article we briefly discussed why having a proper context is essential in the process of knowledge extraction using natural language processing and provided an overview of Parabole's approach and current progress.

About us

Parabole is a Princeton, NJ based cognitive analytics company, automating the creation of enterprise knowledge from unstructured sources of information. We provide our clients with a platform to solve risk, finance and regulatory compliance-related challenges that depend on the knowledge-data interchange. We accomplish this by delivering a range of bespoke applications in the areas of credit, market, liquidity risk and data governance domains.

To learn more, visit www.parabole.ai or reach out to info@parabole.ai