Data Sheet

Datasheet: Causaly Machine-reading Technology Platform

Artur Saudabayev, CTO

Datasheet: Causaly Machine-reading Technology

OUTLINE

The Causaly Machine Reading Platform

Causaly, the provider of a unique evidence-based research platform for Biomedical Cause & Effect discovery, has developed a proprietary technology for understanding Natural Language as published in biomedical literature. It is positioned to transform free text into actionable insights for researchers and decision-makers in Pharma, Life Sciences and Academia.

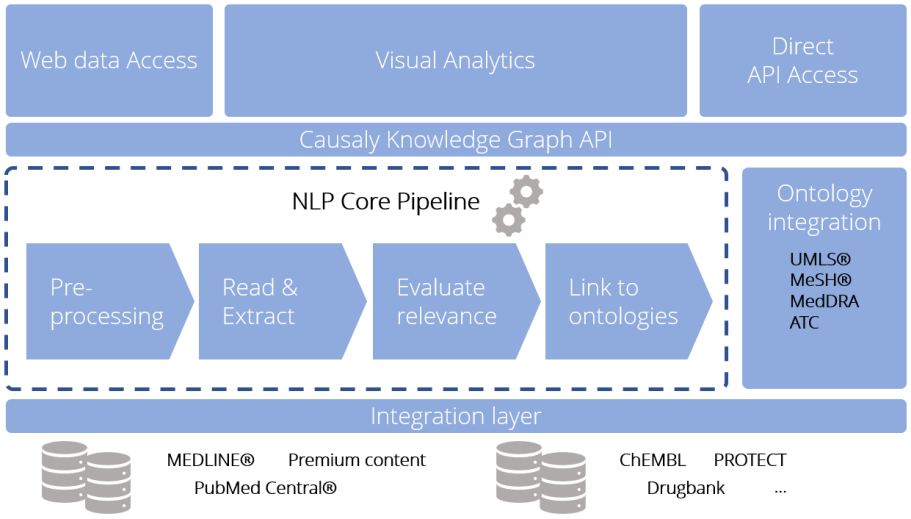

Figure 1 At the core of Causaly is a scalable and easily extensible modular data pipeline

WHAT IS MACHINE-READING ?

Machine-reading is the application of Natural Language Processing Technologies to understand syntax, semantics and discourse of unstructured free text. This includes methodologies for machine learning techniques and symbolic rule-based approaches, with the aim towards Natural Language Understanding.

Causaly Natural Language Processing (NLP) technologies is an ensemble of rule-based linguistic, machine learning & deep learning models trained on proprietary and public data for comprehensive and reliable Event Causality extraction from free text.

The Platform integrates with a variety of sources including MEDLINE™, Pubmed Central™, premium full text content from Publishers as well as 3rd party databases such as ChemBL™, DrugBank™ and others. To date, more than 24,000,000 articles have been processed, yielded approximately 130,000,000 points of evidence on causal and associative entity interactions.

All extracted evidence is traceable to the original research articles and linked to concepts across multiple ontologies for knowledge organization. Ontologies used include UMLS™, MeSH™, MedDRA™ and ATC to name a few, and are continuously extended as required by customers.

All evidence is modelled as a knowledge graph and stored in a graph database that allows computational analysis of connected data. The data is accessible through an Application Programming Interface (API) and a graphical Web-Frontend. Users can use Semantic Search and Boolean Search to find evidence, discover articles and cause-effect relationships and visualize their results without technical knowledge of NLP technologies.

Causaly's modular architecture allows to connect public and customer proprietary data sources, and merge them into one data ocean.

Causaly currently **integrates 7 data sources**, more than **24 million documents** and yields **130 million points of evidence,** growing by 1 million per month.

Flexible data pipeline: Causaly integrates Free text and 3rd party data

The Causaly platform extracts evidence from multiple data sources. It currently integrates the following:

1) **MEDLINE® corpus** of the National Library of Medicine® as currently accessible via PubMed®. More than 24 million English abstract citations are included and updated monthly
2) **PubMed Central® corpus** of the National Library of Medicine® with more than 2 million English full text citations, updated monthly
3) **Premium full text articles** – non-open access English full text citations processed as part of partnerships with publishing groups such as SAGE Publishing.
4) **Structured and curated data sources,** including ChEMBL, Drugbank, RxNorm, PROTECT, DrugCentral and the UMLS Metathesaurus in part or in full.

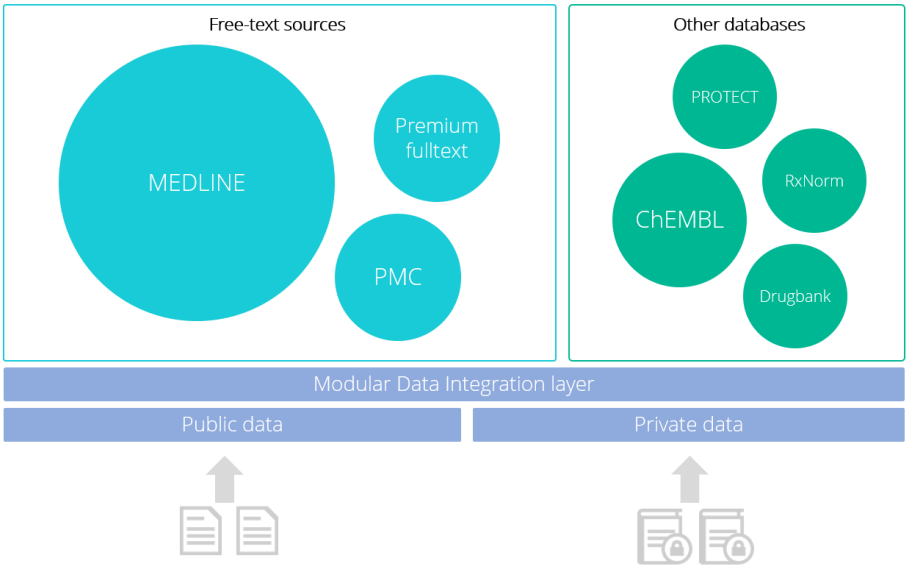

Figure 2 Causaly modular data pipeline integrates public and private data into its knowledge platform

The data platform supports a variety of file formats and is connected through a modular data integration layer with the in-house machine-reading platform. New data sources can be connected to the Causaly platform through preprocessing connectors with minimum effort, and customized to support proprietary data formats.

For more information on integration with your own data sources, please contact us.

At its core, the machine-reading technology **transforms natural language into a structured causal Knowledge Graph, and surfaces evidence from millions of documents.**

The algorithms for extracting complex cause and effect relationships from free text are **proprietary and fully developed inhouse.**

The Core: Causaly Natural Language Understanding platform

Teaching computers how to read and comprehend biomedical publications, as well as extracting cause and effect relationships from them, presents many challenging NLP tasks. But what does it mean to read and understand text for a machine?

"Reading and comprehension" is a multi-step process which broadly refers to the extraction of all the relevant information from a sentence for understanding an affect-relationship or more specifically, Event Causality. This task is concerned with the syntactical and semantic understanding of a sentence: what is the Subject-Predicate-Object agreement, what is the event, where is the action taking place, is it hypothetical or not, is it negated, etc.

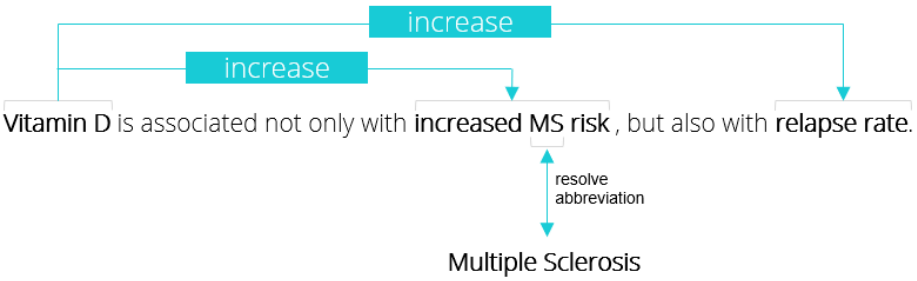

Figure 3 Simple example sentence illustrating part of the machine-reading process.

Subsequently, in-house developed scoring algorithms are used to analyze all evidence for its confidence and linguistic strength values. As a result, evidence can be weak or strong, unidirectional or bidirectional, credible or not.

To further "Understand" evidence, it must be contextualized to define its relevance. For this, factors such as the location of the evidence inside a citation are determined, (for example, Introduction section vs. Results section), the article publication type, the publication year among other factors. The relevance score is later used for sorting results when Users perform searches on the platform.

For the majority of academic papers that lack annotations of publication types, machine learning algorithms for document classification were developed to provide users with predicted document labels such as Randomized Controlled Trials, Case reports and Reviews. This unlocks efficiencies in the literature review process.

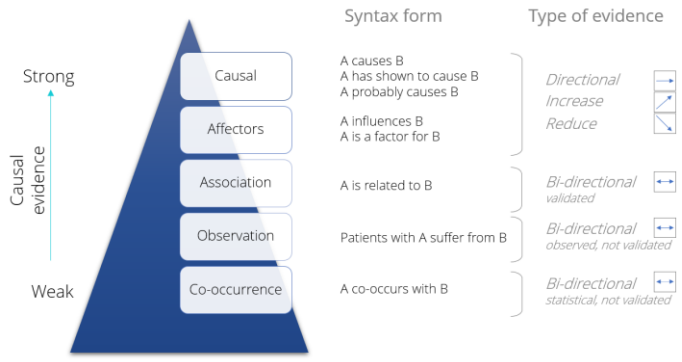

Figure 4 Causaly Knowledge Hierarchy. All evidence is attributed into the hierarchy for scoring

Causaly supports standard ontologies providing interoperability with public and customer knowledge management systems.

New ontologies are integrated on a continuous basis in support of customer requirements, enabling new data insights.

The Use of standard ontologies for better Search and Interoperability

To further organize all knowledge in the graph and to achieve interoperability with other systems and user requirements, all evidence is linked into de-facto standard ontologies. This permits users to access more than 130 million points of evidence through structured dictionaries, and integrate results with existing knowledge management systems.

Causaly currently integrates the following ontologies, and is continuously expanding its vocabularies:

1) Unified Medical Language System® (UMLS®), an ontology that integrates more than 100 vocabularies including MeSH, ICD, SNOMED.
2) Medical Subject Headings® (MeSH®) developed and maintained by the U.S. National Library of Medicine.
3) Medical Dictionary for Regulatory Activities (MedDRA®), a standardized medical terminology to facilitate sharing of regulatory information internationally for medical products used by humans.
4) Anatomical Therapeutic Chemical (ATC) Classification system, used for the classification of active ingredients of drugs, maintained by the World Health Organization.

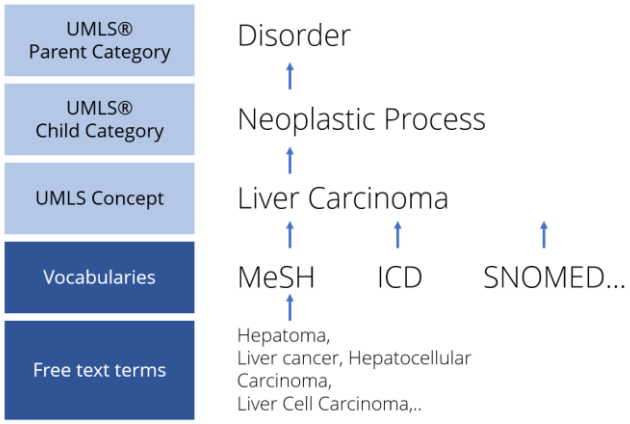

Figure 5 Example how free text terms from biomedical documents are linked into UMLS®

Was this interesting ? Learn more about Causaly machine-reading technology or contact us !

Causaly Inc. is the company behind the evidence-based research platform for biomedical cause and effect discovery. This platform helps researchers and decision makers to find insights among millions of text documents, in minutes.

It is used by Pharmaceutical companies in R&D and Commercial departments

Questions about Causaly ?

Contact us:
info@causaly.com