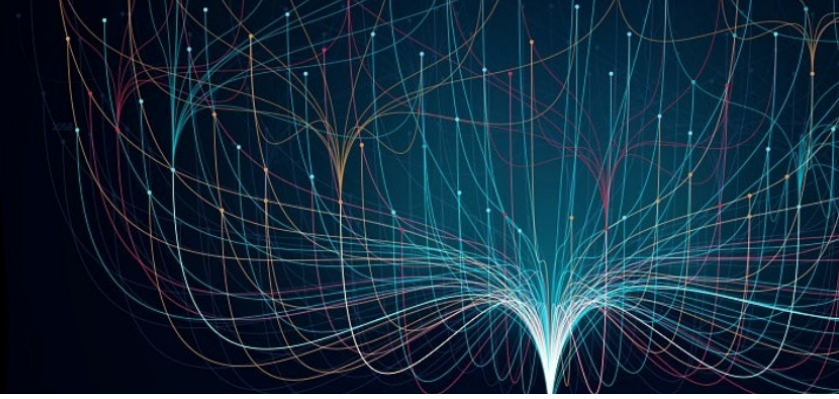
causaly



Accuracy and Precision of Arrow Directionality in the Causaly Knowledge Graph

EXECUTIVE SUMMARY

At Causaly we have developed an advanced machine-reading platform for semantic relationship extraction and comprehension. Our AI system is an ensemble of symbolic and language-based machine-learning models which extracts cause and effect evidence in natural language and uses it to construct directional causal graphs. This form of knowledge representation leverages Causaly's unique AI properties and makes biomedical knowledge computable at scale.

This whitepaper summarizes the purpose, technical approach and accuracy of the arrow directionality resolution algorithm of our machine-reading platform. It describes:

- The problem of knowledge navigation and acquisition from tens of millions of scientific documents. The essence of research of academic literature is in **finding evidence** – isolated statements in natural language hidden among billions of sentences in academic papers.
- How Causaly has developed a powerful, unique machine-reading platform, which is capable of ultra-fast reading of papers and extracting only the relevant evidential statements. Our system then extracts cause and effect relationships, understands their semantics and connects this evidence into a **high-precision knowledge graph** with more than **230 million directional relationships** at the date of publication.
- The high-value ability to express relationships between biomedical concepts in the form of directed arrows, allowing users to **ask complex questions, acquire knowledge quickly and make decisions.** For example, a treatment landscape is expressed as: (drug) – DOWNREGULATE – (disease).
- A reported **precision of 94% for drug-target relationships** and **98% for drug - disease relationships.**

Author
Artur Saudabayev
Co-Founder & CTO, Causaly

For further information,
please visit: www.causaly.com
or email info@causaly.com

INTRODUCTION

One of the primary objectives in biomedical research is to understand how entities at various system levels relate to each other and how they interact. These scientific findings, results of experiments and observations, are reported in natural language and published as academic papers. Our knowledge of how biomedicine works resides in more than 30 million scientific documents.

The large volume of information and the rate of its growth introduces significant challenges in the ability to navigate, explore and discover knowledge. Researchers and domain experts typically look for scientific evidence, which are represented as directional relationships between 2 (or more) biomedical entities (see Figure 1).

To obtain that evidence, however, it is necessary to analyze large sets of documents, firstly by using robust keyword searches, screening and filtering, and a significant amount of time-consuming manual review. At Causaly, we have built a platform that machine-reads and extracts complex relationships from documents. This process enables researchers to find evidence and actionable insights directly, rather than having to review documents first.

At the core of our machine-reading platform is a large Natural Language Processing pipeline, which, in the same way as human-reading, attempts to understand natural language through the lexico-syntactic analysis of sentences and extracting information based on this understanding.

Relationship extraction and biomedical concept identification are critical processes for the extraction of evidence from free text, and the determination of relationship directionality, represented by arrows (see Figure 2) is key to comprehending the nature of the relationship

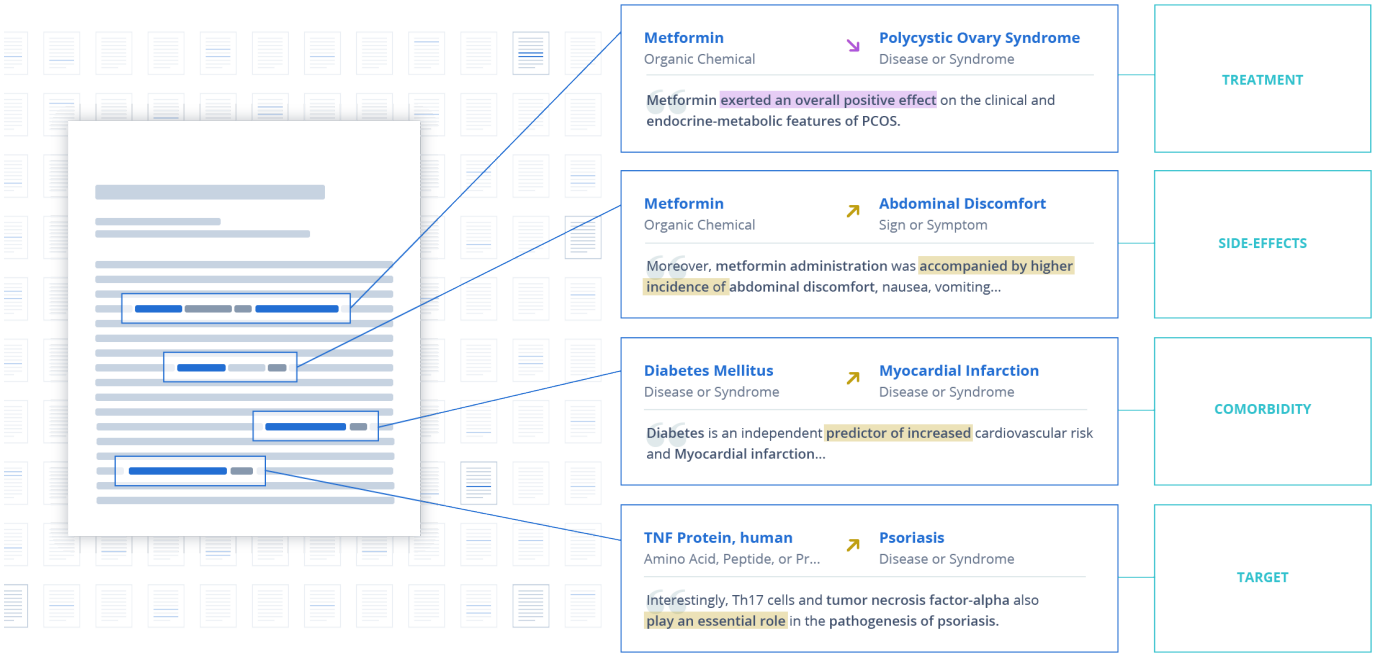


Figure 1. Finding evidence with biomedical research literature

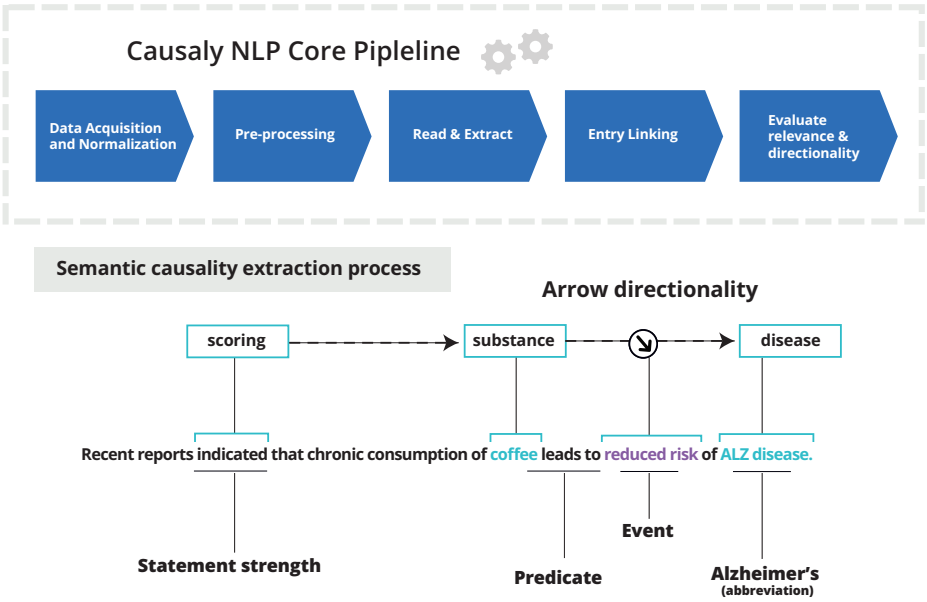


Figure 2. The Causaly Relationship Extraction Pipeline

between two entities, presenting it as knowledge and further making it computable. The following sections will go into detail of how Causaly solves the problem of arrow directionality resolution in complex relationship extraction, a unique differentiating feature of our machine-reading platform, and is supported by associated performance metrics.

THE ROLE OF ARROW DIRECTIONALITY IN BIOMEDICAL EVIDENCE

High accuracy in capturing the arrow directionality of relationships is important for enabling quick and precise decision-making about pressing issues in biomedicine.

The directionality of relationships can reveal important information on how a given substance or gene can be used as a treatment or identified as a cause of side effects. DOWNREGULATE arrows indicate that the given drug is a treatment option for a disease, whereas relationships of an UPREGULATE nature suggest that a disease is a side effect of a drug.

The relationship types shown in Table 1 are particularly important for those who work in the field of clinical research and translational medicine.

Relationship	Description	Biomedical discipline
Drug-Disease	Relationship between drugs and diseases: treatment for a disease or side effect	Clinical research
Gene/ Protein-Disease	Relationship between genes, proteins, enzymes, amino-acids, and receptors and diseases: treatment for a disease or higher chance of/susceptibility for a disease	Translational medicine
Drug-Target	Relationship between drugs and their targets (genes, proteins)	Translational medicine

Table 1. Relationship types and their areas of application

OVERVIEW OF THE ARROW DIRECTIONALITY RESOLUTION MODULE

The nature of the interaction between two concepts, e.g. aspirin (a drug) and headache (a disorder) can be expressed in terms of arrow directionality. Determining arrow directionality is crucial to a correct, computable knowledge representation and semantic interpretation of sentences from biomedical literature.

Causaly's machine-reading platform is a multi-step process consisting of the input (a sentence) being passed through different modules of lexico-syntactic and semantic relationship extraction (illustrated in Figure 3).

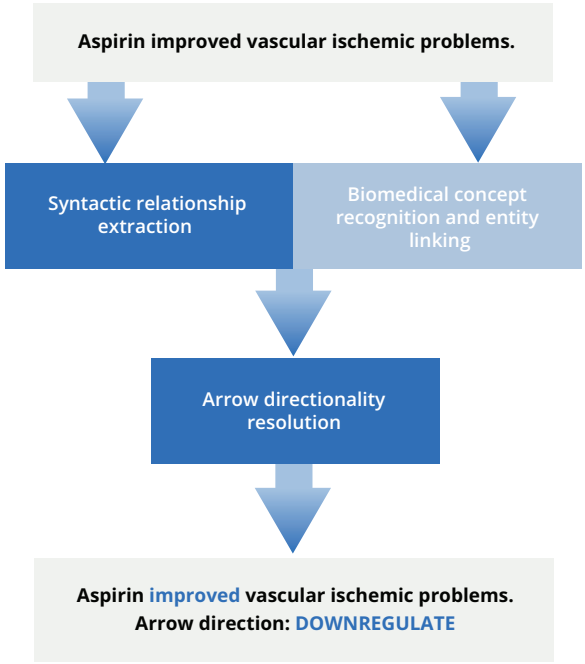


Figure 3. Resolving Arrow Directionality

Lexico-syntactic relationship extraction identifies entities and the linguistic relationship between them (e.g. increase, cause, create, prevent etc.). Subsequently, biomedical concept recognition and entity-linking steps identify these entities as biomedical concepts and locates their semantic types (e.g. Pharmacologic Substance, Disease, Symptom, Food etc.). Knowledge of the linguistic relationship between entities and their semantic types allows us to determine the directionality of the arrow.

ARROW DIRECTIONS

Causaly represents relationships through 8 types of arrows: UPREGULATE, DOWNREGULATE, UNIDIRECTIONAL, BIDIRECTIONAL and 4 of their negative counterparts for refuting statements:

1. UPREGULATE: The cause concept has a positive impact on the effect concept (A increases B) (see Figure 4).

Intracranial hemorrhage was increased by the regular use of aspirin , similarly for both primary and secondary prevention.

This study indicates that aspirin is associated with an increased incidence of intracranial hemorrhage in the author's population.

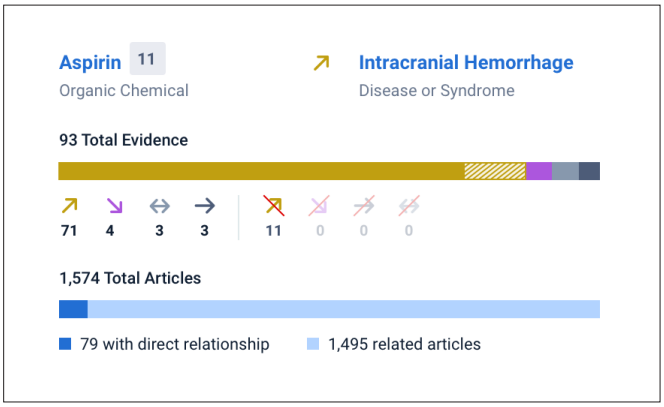


Figure 4. UPREGULATE arrow resolution example

2. DOWNREGULATE: The cause concept has a negative impact on the effect concept (A decreases B) (see Figure 5).

Aspirin significantly reduces the risk of myocardial infarction, stroke, and vascular death in patients with atherosclerotic cardiovascular disease.

We have shown that discontinuation of low dose aspirin increases the risk of non-fatal myocardial infarction in patients with a history of ischaemic events in primary care.

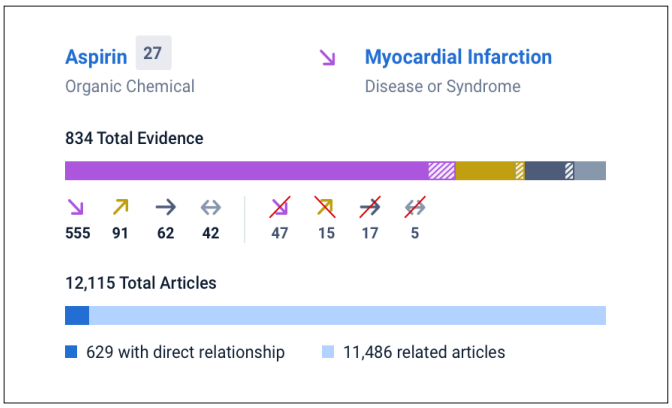


Figure 5. DOWNREGULATE arrow resolution example

3. UNIDIRECTIONAL: The cause concept has an impact on the effect concept, but it is not known whether the impact is positive or negative (see Figure 6).

Aspirin may reduce the decline in cognitive function by influencing multi-infarct dementia , but data are sparse.

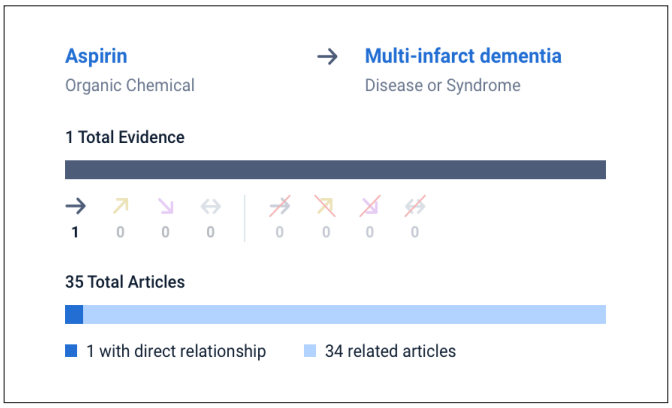


Figure 6. UNIDIRECTIONAL arrow resolution example

4. BIDIRECTIONAL: There is correlation between the cause concept and the effect concept, but it is not known whether the correlation is positive or negative or the directionality is unclear (see Figure 7).

Our present analyses show that mTOR and PTEN expression are associated with the prognosis of ESCC patients, while there are still some notable limitations.

Our immunohistochemical staining data suggest that PKM2 and p-mTOR are both overexpressed in ESCC tissues relative to their expression levels in nontumoral tissues.

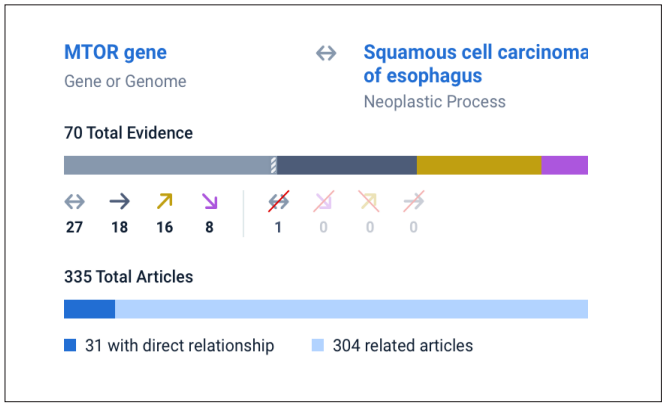


Figure 7. BIDIRECTIONAL arrow resolution example

The correlating refutory statements, such as A does not cause B, A does not reduce B etc. are similarly represented by these 4 arrow types.

ASSESSMENT OF SYSTEM PERFORMANCE

We evaluated the performance of the Causaly system on four annotated baselines: Drug-Disease, Gene-Disease, Drug-Target relationships and a Mixed relationships dataset. Definitions of a “Drug”, “Disease” and “Target” are used here as aggregate terms for selection of UMLS semantic types which constitute the target category (e.g. “Target” is represented by “Gene or Genome” and “Amino Acid, Protein or Peptide” UMLS semantic types).

The datasets were manually annotated by a team of biomedical domain experts and a computational linguist with biomedical expertise. The datasets contained real-life sentences from biomedical publications featured in the Causaly platform. The correctness of arrow directionality was measured for each sentence in the given dataset and overall accuracy was evaluated as the percentage of sentences with a correct arrow direction. For example, if a dataset contains 1,000 sentences, out of which 900 have a correct arrow direction in the system, the accuracy of arrow directionality is 90%.

For aggregated relationships (e.g. the effect of sorafenib on all diseases), accuracy was evaluated as the percentage of relationships with the correct arrow direction. If a data-set contains 1,000 aggregated relationships, out of which 900 are correct, the overall accuracy is 90%. Relationship arrow directions were evaluated by the correctness of the majority arrow direction. For example, out of the 1,394 sentences describing the relationship between sorafenib and liver carcinoma, 956 are of a DOWNREGULATE nature, and therefore the aggregated (majority) arrow direction of this relationship is DOWNREGULATE. As we know that sorafenib is used as a treatment for liver carcinoma, we can tell that the aggregated arrow direction of this relation-ship is correct (Figure 8).

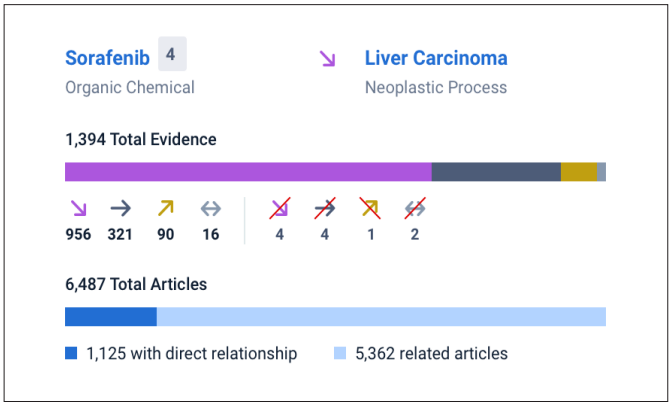


Figure 8. Aggregated arrow direction of the relationship between sorafenib and liver carcinoma

The accuracy calculated for each dataset is shown in Table 2.

Relationship category	Data	Number of evidence	Accuracy
Mixed relationships	Sample from the baseline dataset	1,746 sentences	93%
Drug-disease relationships	Sorafenib-all diseases aggregate data	1,200 aggregate relationships, 10,184 sentences	98% (arrow directionality) 96% (including the correct extraction of concepts)
Gene-disease relationships	Sample from the baseline dataset	2,000 sentences	94%
Drug-target relationships	Adalimumab, secukinumab and infliximab dataset	1,085 sentences	96%

Table 2. Accuracy analysis of different relationship categories