Presentation Abstracts

Welcome address

John Brimacombe, Linguamatics

Mr. Brimacombe will welcome delegates and open proceedings with an update on recent developments at Linguamatics and an overview of our corporate mission - with an emphasis on applications of the I2E platform in both the life sciences & healthcare segments.

Extracting conclusions and interpretations from internal pre-clinical safety reports using I2E

Wendy Cornell, Retired, Merck

Safety assessment studies are a key element of the drug development process and the results of these experiments help determine whether a compound will progress in the pipeline. Much of the data generated in these studies are captured in structured format, however, a significant amount of unstructured information is captured in written reports that describe expert conclusions and interpretations. This unstructured information represents a rich body of knowledge which, in aggregate, has potential to identify capability gaps and evaluate individual findings on active pipeline compounds in the context of broad historical data. We describe the development of a natural language processing (NLP) workflow to extract conclusions and interpretations from Merck's large corpus of internal reports using the Linguamatics I2E software and the integration and analysis of the data using the ANZO platform from Cambridge Semantics.

What's new in I2E in 2015?

Guy Singh, Linguamatics

This presentation provides an overview of the 2 major releases of I2E in 2015, I2E 4.3 and 4.4. As well as providing a summary of all the features within these releases, the presentation will focus on two significant new capabilities introduced to the user community: Connected Data Technology (federated text mining) and EASL (Extraction And Search Language - new query language for I2E). The talk will also include a summary of content released and planned this year on the I2E OnDemand platform.

A systematic examination of gene-disease associations through text-mining approaches

Madhusudan Natarajan, Shire Pharmaceuticals

In 2013, we reported on how we used I2E to tackle repositories of semi-structured information such as the reported patient genotype for disease specific mutations. Deriving systematic annotation around these and correlating these to either efficacy scores, immunogenicity responses etc can offer tremendous insight into patient genotype-phenotype relationships, as well as patient genotype-outcome relationships. We developed a literature corpus around prior reported patient mutations in the public domain and used these to derive a look-up table for a priori prediction of patient disease severity and or response to drug. A meta-analysis of immunogenicity responses to administered drug based on patient genotype was recently presented to and accepted by the European Medicines Agency. We now report how these relationships have been extended to patient registries including to fulfill reporting requirements to regulatory bodies.

Easy access to full text articles for text mining: a new service, a new era

Chris Hilbert, Copyright Clearance Center and Guy Singh, Linguamatics

Researchers struggle to gain access to full text articles for text mining. When they do get the full text they must contend with multiple formats and inconsistent license terms – all of which inhibit text mining efforts. To address these issues, Copyright Clearance Center (CCC), parent company of RightsDirect, has partnered with Linguamatics to make it easier for I2E end users to obtain and index full text XML articles from multiple scientific publishers. This collaboration has led to the introduction of a new service to make high value data sources available for information analysis through text mining. In this presentation we will talk about CCC's new XML for Mining service, the integration with Linguamatics I2E and how the combined solution improves the results of text and data mining queries, reduces costs and mitigates infringement risk.

Full text patent mining: Can it beat manually-curated database subscriptions?

Matthew Crawford, Pfizer

In order to maximize the utility of the information contained in patent applications, Pfizer worked with Linguamatics to employ full-text mining to mine patents relative to diseases of interest. From these patents, the genes targeted by the proposed therapies, the organizations submitting the patent, the overall "invention-type" of the patent, and a relevancy score relating the indication to the patent were derived. The results showed a greater than tenfold increase in the number of patents analyzed with target scoring accuracy approaching that of manually curated patent databases. On top of the raw data, an interface was built to allow scientists to flag interesting targets, comment on the relevance of patents, and track the curation process. The integrated process drastically reduces the FTEs required to keep the organization up-to-date on recent findings published in the patent literature.

Linguamatics I2E in healthcare: recent developments and use cases

Simon Beaulah, Linguamatics

The rapid growth of electronic health records (EHRs) provides an abundant source of valuable data, with the potential to discover insights about patients and their response to treatment. However, with up to 80% of the richest information within the unstructured text, hospitals and medical researchers need better ways to leverage this vital information. Natural Language Processing (NLP) can be used to extract information from the huge range of medical documentation, identifying relevant scientific literature, match patients to clinical trials and power clinical risk models. This talk will look at these and other use cases that enable I2E to add value to EHRs, and how Linguamatics is supporting this vital area.

Extracting data from physicians' notes and surgical path reports using I2E

Samir Courdy, Huntsman Cancer Institute, University of Utah

The amount of structured and unstructured clinical data found in surgical pathology and radiology reports, and physicians notes, including diagnosis, and treatment information is daunting. The effort required for manual abstraction of this information from these reports can be overwhelming. We propose to build an automated workflow process for identifying such reports, utilizing I2E to tag all relevant clinical information, and developing an extraction methodology to associate such unstructured diagnostic information with discrete data elements for research and longitudinal follow up of patients and research subject, to help alleviate the manual and human effort required for abstracting this information, improving quality, consistency and efficiency of data collection for improved outcomes and research.

Sifting through the deluge of information present in surgical pathology reports, physicians' notes, and radiology reports is daunting. To make sense out of all this information, we as informaticists, and data scientists have to develop better approaches and tools utilizing robust data mining techniques and methodology to automatically abstract and annotate data on patients for diagnoses, research cohort identification, and improved outcomes, higher data quality, and reduced costs of manual abstraction. Here, we present a methodology utilizing I2E from Linguamatics as a natural language processing tool for implementing such a solution for prostate cancer, and chronic myelogenous leukemia.

Let Food Be Thy Medicine: Investivating the effects of phytochemicals on human health

Richard Linchangco, University of North Carolina at Charlotte

The prevalence of preventable and chronic diseases, particularly obesity, diabetes, and cancer, has sparked a renewed interest into the effects of diet on human health. Consumption of fruits and vegetables has been linked to reduced risk of cancer and other chronic diseases, but the molecular mechanisms supporting these links remain largely unknown. High throughput (HT) technologies shed light on the components of human diets and their associations with human genetics. The published literature containing this collected knowledge spans many domains of expertise and includes huge volumes of data.
Processing the volume and diversity of scientific literature requires advanced text mining techniques to extract the relationships between dietary components and their molecular interactions with disease related genes.
To investigate the effects of diet on human health, we perform a knowledge-based extraction of semantic predicates between chemicals found in a plant-based diet and genes associated with disease using NLP across large document collections of unstructured-text. Background knowledge is integrated from publicly available resources including NCBI, USDA, and EMBL-EBI. Semantic predicates are validated using data from the UMLS Semantic Network and manual curation.

This study produces a graph-based semantic network of dietary phytochemicals and genes related to disease that assists in the elucidation of health benefits conferred by certain foods. This work provides a foundational resource for further research in nutritional genomics to discover bioactive compounds for disease management through personalized nutrition.

Structuring data using I2E for real world health insights

Jason Evans and Steve Aviv, Pentavere

Healthcare expenditures in developed nations continue to climb as demographics and innovation drive unprecedented utilization and cost. This tsunami of demand has forced governments and insurers to exercise greater control over the management and measurement of health outcomes for its people. In Canada, population health management means increasingly turning to real world health status indicators to better understand the interrelated conditions and factors that influence the health of populations. Innovative healthcare stakeholders, operating in high cost and ultra-competitive markets are seeking out new data sources for insights into consumption, in order to better understand treatment pathways and the value these treatments deliver to those who take them in real world settings.

Pentavere Research Group's "Pentavere Insights RWE" database uses proprietary search programming and Linguamatics I2E to extract data from anonymized health records which provide insights into population health and consumption of health resources.

This presentation will discuss the challenges working with medical records and highlight how I2E is able to breakdown complex document structures and volumes of data in market research studies covering Biologics Medications, Diabetes, Epilepsy, Depression and Chronic Obstructive Pulmonary Disorder.

Future innovations: R&D update

David Milward, Linguamatics

I2E currently provides normalized semantic representations for concepts using either identifiers from existing terminologies or chemical representations such as SMILES or InChi. In this talk we will demonstrate new capabilities in providing normalized representations for concepts which are not defined using fixed terminologies, such as mutations, lab codes, dates, or measurements. Combined with a new capability of range search, this provides a powerful mechanism for finding relevant information however it is expressed e.g. weight over 100lbs whether measured in pounds or kilograms, or articles between 2001 and 2010 across all OnDemand data sources. We will also discuss other developments such as improved processing of conjunctions, incremental indexing and changes to multi queries.

Assessing I2E's applicability for Cancer Research

Uma Mudunuri, National Cancer Institute

Translation of basic biomedical research to a clinical setting requires distilling large amounts of published and internal information, which can be both structured and unstructured, into a focused clinical/biological question. We have evaluated I2E for mining semi-structured text documents and its strengths and weaknesses, using use cases from oncology and other collaborative projects on PTSD and Coagulopathy will be discussed.

Leveraging I2E software for text mining patents on antibody-drug conjugates

Julia Heinrich, Bristol-Myers Squibb

Information in patent publications provides a wealth of opportunities for identifying prior art, licensing deals, and white space. However, the ever increasing number of documents in mostly unstructured format (>95%) brings on the challenges of how to effectively, efficiently and economically extract relevant information in order to make time-sensitive, actionable legal and business decisions. Antibody-drug conjugates (ADC)/immunoconjugates are therapeutic modalities backed by a large and historical portfolio of patents and are fueled to continue to grow by the recent Food and Drug Administration (FDA) approvals of Adcetris and Kadcyla. We have used the I2E text mining software to develop strategies and queries for parsing information from ADC patents into relationship columns in a spreadsheet in order to generate more user-friendly, analysis-ready data outputs. The user case of text mining intellectual property that claim ADC technology for specific antigens will be discussed in this presentation.

I2E in life sciences: recent developments and use cases

Jane Reed, Linguamatics

In this era of big data, life science organizations face the challenge of filtering ever-increasing volumes of text information to gain actionable insights for key decision-making. I2E's flexibility means it can be beneficial in many applications and use cases. This talk will provide an overview of some customer use cases from a range of different disciplines, and highlight some of the solution areas where significant benefit has been found.

Read Alert - Using I2E for automated and centralized alerting of new targets

Jon Hill, Boehringer Ingelheim

Text mining can be an effective solution for retrospective analysis of the literature, but an equally important problem is the quick triage and discovery of new information as it arises. We will describe a system for alerting that uses I2E's capabilities of automation, through scripted indexing and API-based search, and combines them with a database for cross-article comparisons. The resulting system is capable of searching across a variety of primary materials and delivering pertinent results in a timely fashion.

Extracting summary statistics from clinical trial databases

Eric Su, Lilly

Data extraction from clinical literature is necessary for meta-analysis and helpful for designing new clinical trials. Such extraction, usually done manually, is often tedious and error-prone. We utilized I2E to develop automated ways to extract summary statistics on various endpoints from clinical trial databases. This presentation will introduce some basic queries that are designed to extract such data. Two databases, TrialTrove and ClinicalTrials.gov, and two therapeutic areas (oncology and diabetes) will be used in the examples. These queries can be deployed through "Smart Query" or Web GUI so that others, including non-I2E users, can edit query parameters and generate user-defined output in html or Excel format.

Extracting and applying information from full text scientific articles for patient care

Jonathan Hartmann, Georgetown University Medical Center and Guy Singh, Linguamatics

The past couple of years has seen Georgetown University Medical Center (GUMC) make innovative use of information extracted from medical abstracts using I2E. Providing this information during daily clinical hospital rounds at the patients bedside has meant that physicians have had immediate access to the vast amount of information in MEDLINE. This talk looks at how this has now been extended to extracting information from full text articles in Elsevier ScienceDirect. It will look at how the additional information provided in these journals has been used by clinical staff through the use of real life anonymized example cases. The presentation will include a description of how this solution evolved and the various hurdles that needed to be overcome from both a legal and technical angle.