Find the signal in the noise
The adoption of Electronic Health Records (EHRs) in the USA is rapidly increasing, due to the Health Information Technology and Clinical Health Act (HITECH).

EHRs create a vast source of medical data. Hidden within them is the promise of improved healthcare outcomes, disease understanding and drug development. But EHRs are difficult to exploit.

EHR datasets are complex. Around 75-80% of data in EHRs is unstructured and stored as text fields, making it hard to extract insights from patient documentation.

Significant data and insights are trapped in the unstructured fields – information that is essential for predictive modelling, population modelling, clinical research and analytics.

How can we unlock this data?

---

1. Frost & Sullivan, Electronic Health Record Usability: CIOs Weigh In, Oct 2014

"Natural language processing (NLP) and visualization dashboards are the technologies most suitable to improve EHR usability. NLP can produce readable summaries of unstructured text, helping clinicians retrieve information needed for point-of-care decision making.” 1
Extracting the value from unstructured data

To make effective use of unstructured data in a clinical domain, a system must be able to distinguish concepts and their context within the text. It is essential to differentiate current diagnosis from family history of disease, when a disease term has been negated or ruled out and identify what drugs and dosages have been used.

This is the sort of task that Natural Language Processing (NLP) can address. NLP allows more precise and efficient extraction of knowledge by capturing the different ways people express the same information.

Drawing from data across hospital departments and systems, NLP can support medical researchers, data analysts and clinicians in areas such as:

- Extracting patient data for use in data warehouses, biobanks and disease registries
- Cohort selection for clinical trials and translational studies
- Predictive modelling of disease, treatment and patient outcomes
- Correlation between wellness and patient adherence to care pathways
- Alerting of potential adverse reactions
- Assessment of case histories of disease co-morbidity and treatment

Figure 1  The ecosystem of organizations that surround Electronic Health Records
New frontiers

As healthcare customers come to understand the capabilities of NLP, they see a new world of possibilities emerging.

Imagine joining data from an EHR with contraindications in FDA drug labels to alert for potential adverse events; or mining ClinicalTrials.gov’s inclusion and exclusion criteria to find a trial for a patient.

This marriage between structured, semi-structured and unstructured data, and the synthesis of knowledge across multiple information silos – is already enabling healthcare to move to a more efficient, evidence-driven approach.

The next generation of Healthcare NLP

Medical use of NLP is well-established. It is necessary because pathology, laboratory and radiology reports, and doctors’ notes are often dictated and transcribed.

But truly valuable insights can only come with deep linguistic analysis: understanding of context, negation of terms and the morphology of language. All of which has to be combined with knowledge about medical concepts – the drugs, diseases, measurements and techniques that are used.

And this insight is what the comprehensive NLP capabilities of I2E bring to the field.

Introducing Linguamatics I2E

Linguamatics’ agile NLP solution, I2E, supports all of these complex tasks, and is easy and intuitive to use.

I2E takes complex and largely unstructured documents such as pathology and radiology reports to produce structured data sets systematically.

These data sets can then be used to build warehouses of research data, or provide analytics to support patient care – for example, in the retrospective analysis of a particular treatment.

I2E can also be used to augment search. Data can be enriched, or clustered semantically, or tagged by concept for easier navigation.

Key benefits

◆ Extract patient insights from multiple unstructured records
◆ Provide detailed structured data for use in analytics, data warehouses, predictive modelling and cohort selection
◆ Support comparative effectiveness and outcomes analysis
◆ Provide fast, iterative query building giving rapid time to value
◆ Enable non-developers to use NLP, removing bottlenecks
◆ Support linguistic wildcards to explore large data sets, reducing the reliance on gold standard annotations
I2E can be used to

- Break down sentences into meaningful units
- Identify and extract concepts based on domain ontologies
- Assign terms to appropriate concepts
- Identify the relationships between concepts
- Recognize when terms are in a negative context
- Handle grammatical variation, to standardize different constructions

I2E's agile and powerful user interface allows non-developers to quickly build knowledge extraction queries - unlike traditional NLP approaches, where rules have to be coded into the system by specialists, or require extensive clinician-annotated training data.

The flexibility and openness of I2E also contrasts with the more rigid use of NLP in Computer Aided Coding systems.

Deep linguistic analysis

I2E uses deep linguistic analysis of text to deliver accurate and focused sets of results, which then reveal clearer insights.

Here are some examples:

- **Identify subtleties of language** - I2E understands the difference between “history of cancer” and “family history of cancer”.

- **Identify numerical information in context** - such as dosages of specific drugs or tumor size.

- **Manage negation in medical text** - by flagging concepts based on negative terms and the linguistic context. Thus, “No evidence of pneumonia” does not return a diagnosis of “pneumonia”.

- **Identify, cluster and categorize by different concepts and classes** - using plugged-in terminologies, thesauri or ontologies. For example: breast cancer and all its synonyms; or any type of cancer.

- **Use any terminologies** - such as ICD-9 & 10, RxNorm or SNOMED CT.

- **Connect and merge the I2E results table with other structured data** - whether from a relational database or from other I2E results.
Using I2E with healthcare data

I2E takes its input from common EHRs such as Epic, Cerner and AllScripts, as well as department level solutions such as Copath.

These reports and text fields are combined with domain knowledge such as ontologies and coding standards to provide sophisticated, highly scalable indexes.

I2E then provides multiple querying options, from entity extraction through to full linguistic analysis. Small queries can be combined to build a full patient profile; smart queries can be deployed to a wider audience based on more sophisticated searches created by power users.

The agility and speed of I2E is what really sets it apart. It allows an iterative, data-driven approach to query building that takes seconds to complete. All very different from the drawn out training processes usually associated with NLP.
Case studies

1. Pathology report extraction

Huntsman Cancer Institute (HCI) had long recognized that valuable research insights were hidden in pathology documents. Things like tumor size, type and stage; and lab results showing bone marrow and blood measurements, when properly analyzed, provide much richer data.

Together, Linguamatics and HCI work to mine cancer pathology reports systematically, extracting relevant data and loading them into their research data warehouse.

In due course this data will support cohort selection for all types of clinical and translational studies, providing insight into different cancer types.

For more information, please see our Application Note: Accelerating Information Extraction from Clinical Records.

2. Pneumonia prediction

Linguamatics have collaborated with Kaiser Permanente to develop a system that categorizes potential pneumonia patients based on interpretation of their chest X-rays in radiology reports.

Many subtleties of language are used by radiologists when writing their notes and the diagnosis is generally inferred from a set of individual observations. In the past, this inference was performed manually, making it difficult to use radiology reports at any scale in retrospective analyses.

Now, however, physician-developed criteria are used by I2E to categorize the reports automatically into the following categories:

1. The patient has pneumonia
2. The patient possibly has pneumonia
3. The patient doesn’t have pneumonia

The system was assessed against a 300 patient gold standard and achieved selectivity and specificity scores of over 90%. It has since been run with over 200,000 patient reports.

To fully evaluate the unique and compelling benefits that I2E can bring to your organization, please contact your local Linguamatics representative or email us at enquiries@linguamatics.com.

About Linguamatics

Linguamatics is the world leader in deploying innovative, health-science-focused NLP solutions for high-value knowledge discovery, information extraction and decision support.

Linguamatics I2E is used by leading hospitals, the US Food and Drug Administration, premier academic institutions and 17 of the top 20 global pharmaceutical companies. They use it to take unstructured and semi-structured ‘Big Data’ and deliver improved patient care and insights.

I2E mines a variety of text resources, such as electronic health records, pathology and radiology reports, initial assessments, discharge summaries and ICU notes to support knowledge extraction for use in data warehouses, disease registries, predictive models and healthcare analytics.

I2E allows medical information experts who are not developers to quickly build and modify queries that extract information which is critical to supporting Meaningful Use, comparative effectiveness and adherence to care pathways.

Linguamatics is committed to excellence in healthcare informatics and is a corporate member of AMIA and HIMSS. The company operates globally, with headquarters in Cambridge, UK, and a U.S. office near Boston, MA.

About I2E

I2E is an agile, scalable, high performance text mining system that facilitates discovery and knowledge synthesis from unstructured text in large document collections.

I2E is able to mine any text-based resources, such as electronic health records, pathology and radiology reports, initial assessments, discharge summaries and ICU notes to support knowledge extraction for use in data warehouses, biobanks, disease registries, predictive models and healthcare analytics. I2E allows medical information experts who are not developers to quickly build and modify queries that extract information which is critical to supporting Meaningful Use, comparative effectiveness and adherence to care pathways.

There is a choice of ways in which you can connect to I2E’s unique capabilities: either by deploying I2E Enterprise in-house, or via I2E OnDemand, our Software-as-a-Service (SaaS) version of I2E.

For more information, visit www.linguamatics.com or www.linguamaticshealth.com