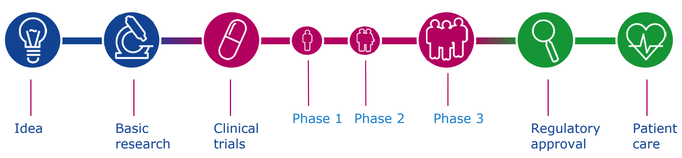
Bridging the gaps: NLP in translational research

Jane Z Reed, Head of Life Science Strategy, Linguamatics

NLP for Biomedicine

- For the past few years, artificial intelligence (AI) technologies such as natural language processing (NLP) have been hot topics in biomedicine, as researchers and healthcare providers consider ways to leverage innovative tools to transform bioscience research and clinical care, from bench to bedside.
- Accessing the right information is critical but much of the data is locked in textual format, such as scientific literature, clinical trial reports or electronic health records. NLP can effectively speed the extraction of critical information from unstructured scientific and clinical text.
- Use cases span discovery, development, and healthcare delivery – such as utilization of text mining for genotype-phenotype annotation, selecting patients for clinical trials, and extracting key endpoints from pathology reports and EHRs for better patient care.

Accessing the right information is critical for translational research, across drug discovery, development and into delivery of healthcare therapeutics



NLP enables deep linguistic analysis

- I2E uses deep linguistic analysis of text to deliver accurate and focused sets of results, which then reveal clearer insights. For example:
  - **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 tumour 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.

NLP Text Mining in Biomedicine

- To make effective use of unstructured data in the biomedical domain, we need 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.
- NLP can do this. MLP allows more precise and efficient extraction of knowledge by capturing the different ways people express the same information.
- NLP: Find information however it is expressed, and understand relevant context

Different words, grammar, same meaning
5mg/kg of cyclosporine per day
5mg/kg per diem of ciclosporin
Neoral 5mg per kg per day

Different expression, same meaning
Non-smoker
Does not smoke
Does not drink or smoke
Denies tobacco use

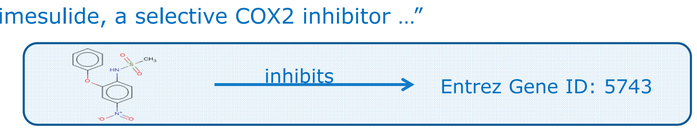
Same word, different context
Diagnosed with diabetes
Family history of diabetes
No family history of diabetes

Normalisation of data

- Finding concepts is not enough for rapid analysis and actionable information. Use of ontologies and pattern rules enables **normalisation** of concepts, for semantic tagging and data integration.

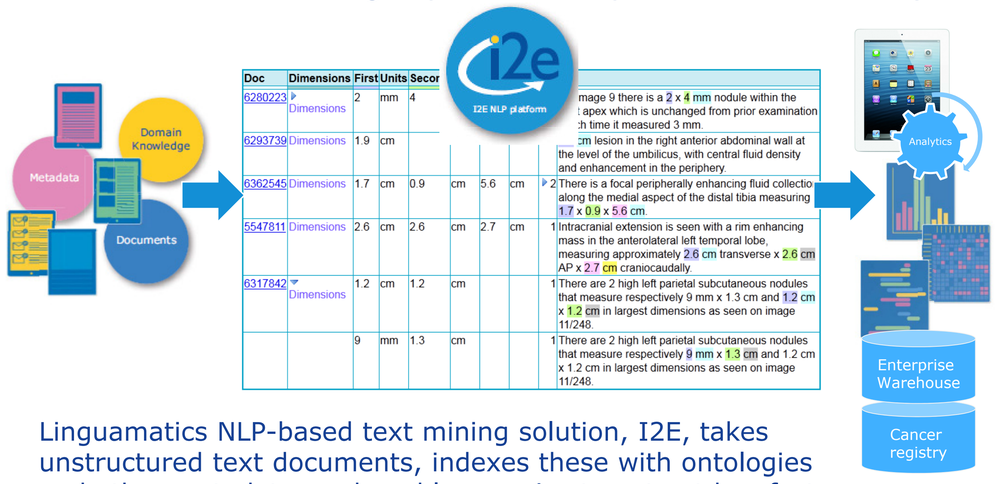
Concept	Text	Normalized Value
Diseases	breast cancer	Breast Neoplasm
	carcinoma of the breast	
Genes	Raf-1	RAF1
	Raf I	
Dates	27th Feb 2014	20140227
	2014/02/27	
Measurements	0.2g	200 mg
	Two hundred milligrams	
Mutations	Val 158 Met	V158M
	Val by Met at codon 158	

"Nimesulide, a selective COX2 inhibitor …"



I2E Transforms Text into Actionable Insights

Turn text …into structured data using sophisticated queries …to drive analytics



Doc	Dimensions	First	Units	Secor					
6280223	Dimensions	2	mm	4					mage 9 there is a 2 x 4 mm nodule within the apex which is unchanged from prior examination h time it measured 3 mm.
6293739	Dimensions	1.9	cm						cm lesion in the right anterior abdominal wall at the level of the umbilicus, with central fluid density and enhancement in the periphery.
6362545	Dimensions	1.7	cm	0.9	cm	5.6	cm	2	There is a focal peripherally enhancing fluid collectio along the medial aspect of the distal tibia measuring 1.7 x 0.9 x 5.6 cm.
5547811	Dimensions	2.6	cm	2.6	cm	2.7	cm	1	Intracranial extension is seen with a rim enhancing mass in the anterolateral left temporal lobe, measuring approximately 2.6 cm transverse x 2.6 cm AP x 2.7 cm craniocaudally.
6317842	Dimensions	1.2	cm	1.2	cm			1	There are 2 high left parietal subcutaneous nodules that measure respectively 9 mm x 1.3 cm and 1.2 cm x 1.2 cm in largest dimensions as seen on image 11/248.
		9	mm	1.3	cm			1	There are 2 high left parietal subcutaneous nodules that measure respectively 9 mm x 1.3 cm and 1.2 cm x 1.2 cm in largest dimensions as seen on image 11/248.

Linguamatics NLP-based text mining solution, I2E, takes unstructured text documents, indexes these with ontologies and other metadata, and enables queries to extract key facts in a structured form, that can then power analytics workflows e.g. in clinical risk models.

TNM Staging from Pathology Reports

Cancer information is one of the most common areas for NLP to be applied to in healthcare, providing key insights into disease progression and long term outcomes. Valuable research insights hidden in **pathology documents** can be extracted, such as **tumour size, type and stage**; and lab results showing bone marrow and blood measurements, providing much richer data for diagnostics and risk prediction.

I2E is used in healthcare organisations to mine cancer pathology reports systematically, extracting and structuring relevant data e.g. tumour, node and metastasis (TNM) staging.

TNM Stage	Doc	Hit
T2 N0 M0	5 PMC2955275/3	1 ... for invasive ductal cancer (T2N0M0, stage IIa).
pT3 N0 M0	2 PMC2621222/2	1 ... stage of the tumor was pT3N0M0 R0.
T3 N0 M0	2 PMC3337737/3	1 ... study reported an adenocarcinoma, T3N0M0.
pT1b N0 Mx	1 PMC3024927/2	1 ... ovarian papillary serous adenocarcinoma (pT1bN0Mx).
pT2 N0	1 PMC3016302/3	1 ... International Cancer Control classification was pT2, pN0 (0/12)...
pT3 N1 M0	1 PMC2621222/1	1 ... a rectal adenocarcinoma, staged pT3 N1 M0 R0 with complete mesorectal excision ...
pT3 N2 M0	1 PMC2843670/2	1 ... lymph node metastasis (Stage pT3N2M0).
T1 N3 M0	1 PMC3010544/4	1 ... diagnosed as hypopharyngeal cancer (T1N3M0) underwent 5-weeks of radiation ...
T1b N0 M0	1 PMC3010544/3	1 ... diagnosed as glottic cancer (T1b-N0M0) underwent 8-weeks of radiation ...
T4 N1 M0	1 PMC3276824/1	1 ... carcinoma supraglottic larynx (stage T4N1M0).

Stage Type	T Stage	N Stage	M Stage	Doc	Hit
Clinical	T2	N0	M0	6 PMC2955275/3	1 ... for invasive ductal cancer (T2N0M0, stage IIa).
Pathological	T3	N0	M0	2 PMC2621222/2	1 ... stage of the tumor was pT3N0M0 R0.
	T2	N0		2 PMC2194707/1	1 TNM for PTC was pT2, N0 and for MTC was stage ...
	T4	N1	Mx	1 PMC2761909/2	1 ... confirmed a Dukes C1 (pT4N1Mx) mucinous adenocarcinoma of the ...
	T1	N0	Mx	1 PMC3024927/2	1 ... ovarian papillary serous adenocarcinoma (pT1bN0Mx).
	T3	N1	M0	1 PMC3218132/2	1 The final pathologic stage was T3N1M0, stage IIIB.

Structured tabular results from I2E NLP queries to extract tumour staging information from full text pathology reports. I2E's flexibility enables outputs to be formatted to accommodate the needs of the user or organisation.

Genotype-Phenotype in Rare disease

- Shire Pharmaceutical's innovative implant device can deliver enzyme replacement therapy to the CNS of **Hunter Syndrome** patients. The procedure is invasive and unpleasant, so ideally would be provided to those infants likely to have more **severe phenotypes**.
- Shire used I2E to identify all the iduronate-2-sulfatase (IDS) **gene mutations** recorded in Hunter patients (and patients with Hunter-like symptoms) and link them to phenotypes, enabling a clear picture of which genotypes were associated with the more **severe cognitive impairment phenotypes**.

Shire

- Increased understanding of IDS mutational spectrum for **provider diagnostics**
- Enabled **better clinical decision-making** and a focused precision medicine approach for patient care

PMID	Source	Mutation Genes/Prote..	Severity	Doc	Hit
8111411	PDF	Q531X	mild	2 Hopwood_gene_8111411	1 ... and R48P, L196S, Q531X (mild phenotype)
15614569	PDF	H138R	severe	2 Chang Ex__II_15614569	1 Patients with R88C and H138R mutations displayed a severe phenotype
17391447	PDF	E177X	attenuated	2 Froissar_ppl_17391447	1 In contrast, the attenuated phenotype reported in the patient carrying the E177X mutation (26) is ...
9660053	PDF	nonsense mutation	very mild	2 Froissar_enet_9660053	1 This nonsense mutation is associated with a very mild phenotype (patient 56, aged ...
24125893	PDF	c.1122C>T	attenuated	1 Mucopoly_nts_24125893	1 ... mutations correlation with the attenuated form (c.1122C>T), while a greater ...
24780617	Abstract	p.Ile360Tyrfs*31	severe	1 24780617	2 ... mutations whereas the p.Ser142Phe and p.Ile360Tyrfs*31 mutations caused the severe disease manifestation
9712538	PDF	A deletion involving exons 2-4 in the iduronate-2-sulfatase gene	intermediate	1 Bonuccel_enet_9712538	2 A deletion involving exons 2-4 in the iduronate-2-sulfatase gene of a patient with intermediate Hunter syndrome
1284597	Abstract	R468W	mild	1 1284597	1 Mutation R468W of the iduronate-2-sulfatase gene in mild Hunter syndrome (mucopolysaccharidosis type II) ...
7887413	PDF	P469H	mild	1 Jonsson_enet_7887413	1 ... mutations in exon 9 had mild disease (P469H, Y522C, R469W, ...
7981716	PDF	R468W	mild	2 Mutation_S II_7981716	1 ... C (1992) Mutation R468W of the iduronate-2-sulfatase gene in mild Hunter syndrome (mucopolysaccaridosis type II) ...
8566953	Abstract	A346D	mild	1 8566953	1 The A346D mutation was associated with the mild phenotype, all others with the ...
9501270	PDF	Q389X	severe	1 Isogai_1_bDis_9501270	1 ... nonsense mutations (Q92X, Q389X) in patients with severe Hunter syndrome (mucopolysaccharidosis type III).

NLP for Pneumonia prediction

Kaiser Permanente used I2E 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

KAISER PERMANENTE®

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.

Liu et al. BMC Med Inform Decis Mak. 2013 Aug 15;13:90
PMID: 23947340

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 healthcare organisations (providers, medical centres, payors), the US FDA, and 18 of the top 20 global pharmaceutical companies. They use I2E to take unstructured and semi-structured 'Big Data' and deliver improved patient care and insights.

For more information, visit www.linguamatics.com or contact info@linguamatics.com