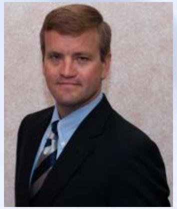
NLP SUMMIT HEALTHCARE

Extracting Big Value From Big Data in Unstructured Content



Paul Edelblut

Vice President

Vantage Labs

Who We Are

Headquarters in USA with operations worldwide

Leaders in advanced AI, Big Data solutions, Natural Language Understanding

2.2 Billion daily customers (likely including you!)

100 million text assessments annually

Operational projects since 1997

Serving healthcare and adjacent industries for decades

Paul Edelblut, VP of Global Operations for more than 20 years

Some of our Customers

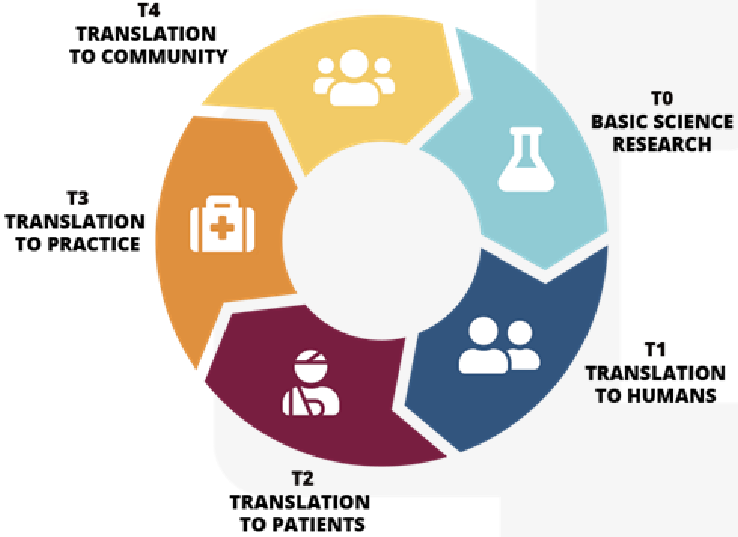
Perelman
SCHOOL OF MEDICINE
UNIVERSITY of PENNSYLVANIA
COLUMBIA
VAGELOS COLLEGE OF PHYSICIANS AND SURGEONS
CASE WESTERN RESERVE UNIVERSITY
EST 1826
SCHOOL OF MEDICINE
M
UNIVERSITY OF MICHIGAN MEDICAL SCHOOL
MICHIGAN MEDICINE
AMERICAN COLLEGE OF SURGEONS
Inspiring Quality:
Highest Standards, Better Outcomes
100+years
Mc Graw Hill Education
CQUniversity
AUSTRALIA
The Royal Australian and New Zealand College of Radiologists®
ROSALIND FRANKLIN UNIVERSITY
of MEDICINE AND SCIENCE
EST 1912
TuftsMedicine
Microsoft
THE WARREN ALPERT
Medical School

Why Vantage:

- ➢ 20 years of proven implementation experience
- ➢ Fusion of wide range of data through NLU
- ➢ Unifying disparate systems
- ➢ Exposure/experience across many sub-specialty domains
- ➢ Multi-national implementations for linguistic variety
- ➢ Decades of experience

The Challenge and Opportunity

- Demand for more translational research
- Default search of less than authoritative resources
- Rich content available
- First Step is unifying data and making it accessible

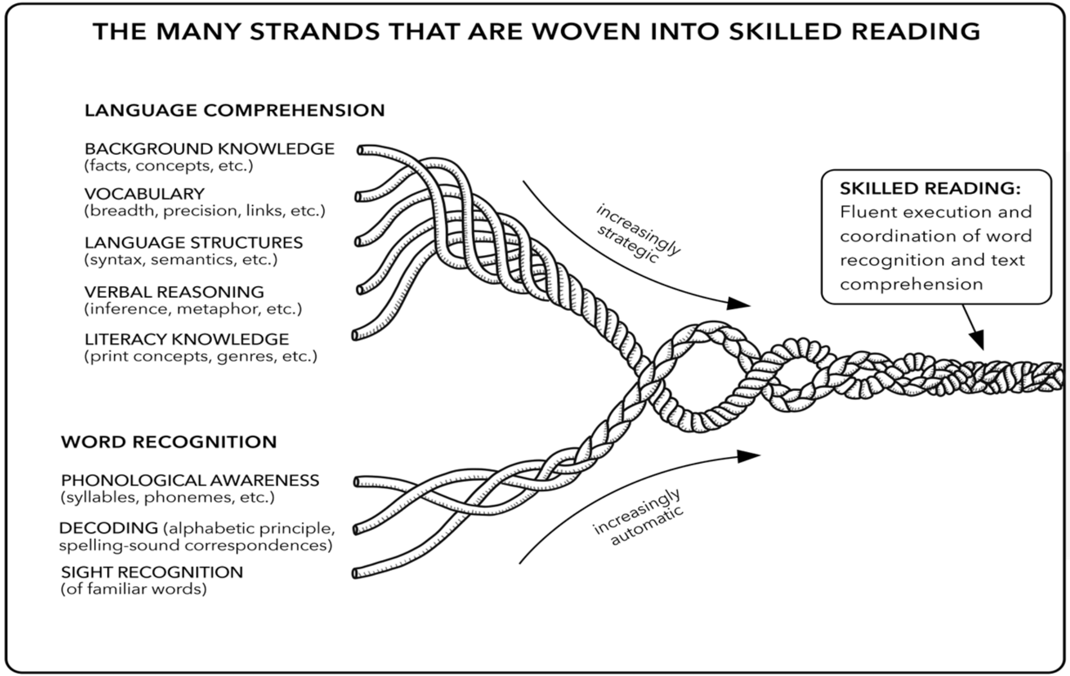


©URMC

NLP SUMMIT

VANTAGE

A Neurosynthetic Approach



THE MANY STRANDS THAT ARE WOVEN INTO SKILLED READING

LANGUAGE COMPREHENSION

BACKGROUND KNOWLEDGE
(facts, concepts, etc.)

VOCABULARY
(breadth, precision, links, etc.)

LANGUAGE STRUCTURES
(syntax, semantics, etc.)

VERBAL REASONING
(inference, metaphor, etc.)

LITERACY KNOWLEDGE
(print concepts, genres, etc.)

WORD RECOGNITION

PHONOLOGICAL AWARENESS
(syllables, phonemes, etc.)

DECODING (alphabetic principle,
spelling-sound correspondences)

SIGHT RECOGNITION
(of familiar words)

SKILLED READING: Fluent execution and coordination of word recognition and text comprehension

Identifying paraphrases with deep neural networks

How it works...

Consider the following two sentences:

I suspect pancreatic fibrosis.
It could be a fibrocystic disease of the pancreas.

Step 1: Encoding text in machine-readable form for vectors and tokens

fibrosis: [1.7657, 1.0454, 0.35474, ..., -0.52842, -0.16688, -0.15263, 0.33342]

Step 2: Feed encoded text into the network

Step 3: Interpreting the result as 1 or 0 corresponding to paraphrase or non-paraphrase

- engine is trained on an accurate and representative dataset. In essence, we ask our classifier to learn a function *f*, such that
    - $f(\cdot) : R^m \to R^o$
- where *m* is the dimension of our input vector and *o* is the dimension of our output vector (in this case, 2). Note, the input dimension will be determined by (a) the dimensionality of our word embeddings, and (b) any additional vectors or padding added to those initial embeddings within the network.

Parsing the query

[you [**pronoun**]**NH]** |**[** link [**verb**]VH**]** | **[**a [**article**]**N** | microsoft [**proper** name]**N** | excel [**proper** **name**]**N** | spreadsheet [**noun**]**NH]** |**[** to [**preposition** **PH**] |[a [**article**]NP | part [**noun**]**NP** | file [**noun**]NH]**P]** | but [**conjunction**] |**[** the [**article**]**N** | parameter [**noun**]**NH]** |**[** be [**verb**]**V** | not [**adverb**]**V** | display [**verb**]**VH]**

Phase 1: Tokenization

Sentences are broken up into tokens, i.e. words and punctuation signs.

Phase 2: Morphological Analysis

Tokens (other than punctuation signs) are looked up in a morphological dictionary, and parts of speech and attributes are assigned. Inflected forms are rooted to their base forms. Some tokens will have more than one reading like e.g. “are”.

Phase 3: Disambiguation

Ambiguous tokens (i.e. tokens with more than one part of speech) are disambiguated based on the morphological and syntactical context. For example: the token “link” is a verb, since it is preceded by a personal pronoun. The token “to” is a preposition, since it is followed by an article and a possible noun compound. The token “but” is likely to be a conjunction in this sentence, since it is preceded by a comma.

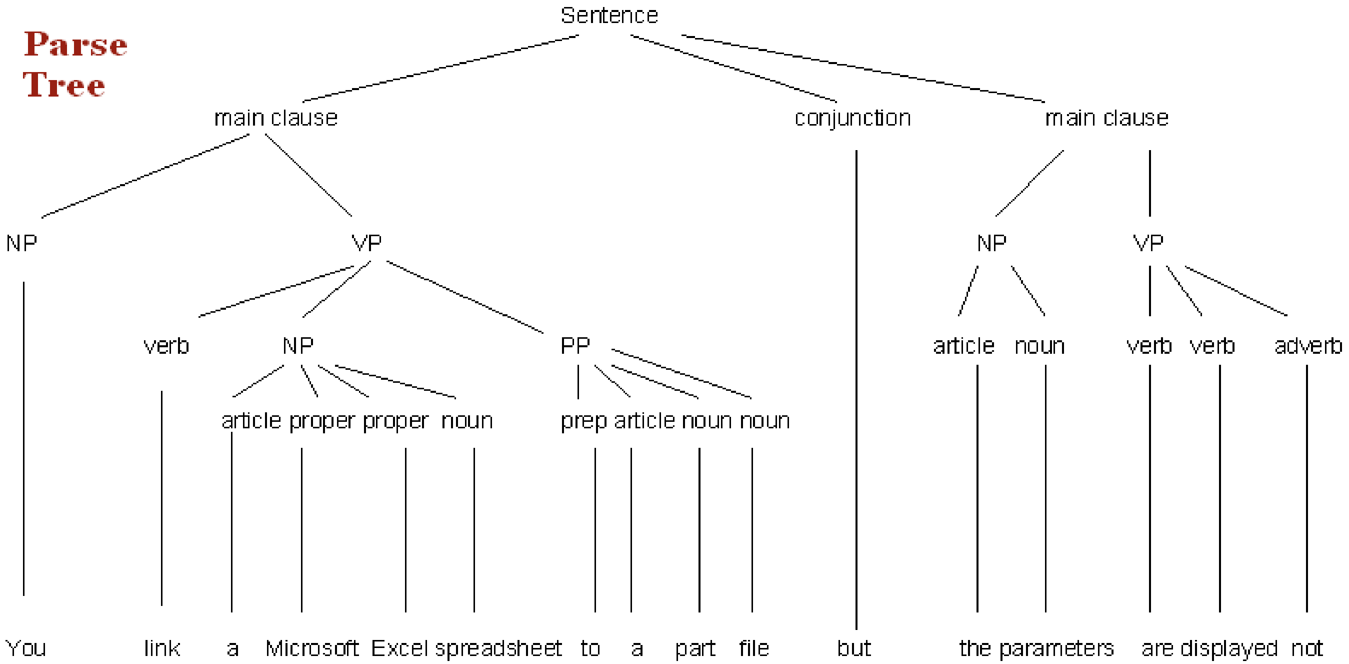
Phase 4: Phrase boundaries

In a next step, phrases are formed; noun phrases [N], verb phrases [V] and prepositional phrases [P]. The main words in each phrase are called phrase heads: NH, VH, PH. The head of a noun phrase is e.g. a noun (“spreadsheet”), or a pronoun (“you”).

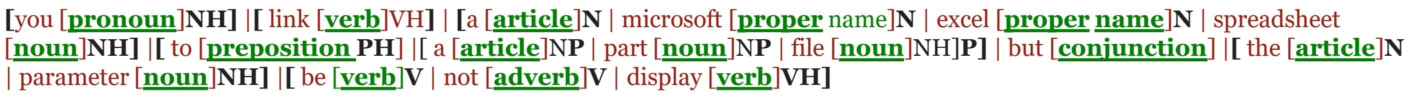
Parsing the query

[you [pronoun]NH] |[link [verb]VH] | [a [article]N | microsoft [proper name]N | excel [proper name]N | spreadsheet [noun]NH] |[to [preposition PH] |[a [article]NP | part [noun]NP | file [noun]NH]P] | but [conjunction] |[the [article]N | parameter [noun]NH] |[be [verb]V | not [adverb]V | display [verb]VH]

Parse Tree



Normalizing the query



[you [pronoun]NH] |[link [verb]VH] | [a [article]N | microsoft [proper name]N | excel [proper name]N | spreadsheet [noun]NH] |[to [preposition PH] |[a [article]NP | part [noun]NP | file [noun]NH]P] | but [conjunction] |[the [article]N | parameter [noun]NH] |[be [verb]V | not [adverb]V | display [verb]VH]

- Rooting
    - parameters ⇒ parameter
    - displayed ⇒ display
    - linking ⇒ link
    - parts ⇒ part
    - you, I ⇒ *pp*
    - roots are marked *
- Preserving phrase boundaries and giving full phrase matches a high weight (|*part* *file*|)
- Strong synonym normalization to its canonical form (MS ⇒ Microsoft)
- Weaker synonym for query expansion (connect=link)
- Language Variety (colour ⇒ color)

Building the linguistic image

- ## The sentence:

1. You link a Microsoft Excel spreadsheet to a part file, but the parameters are not displayed.
    - *pp* *link* |Microsoft Excel *spreadsheet*| to |*part* *file*| but *parameter* *no* *display*

- ## Question variations:

2. I connected a MS spreadsheet to a parts file and no params get displayed?
    - *pp* *connect=link* |Microsoft Excel *spreadsheet*| to |*part* *file*| and *no* *parameter* get display*
3. Having a problem linking a spreadsheet to the parts file.
    - *have* *problem* *link* *spreadsheet* to |*part* *file*
4. How can one get the parameters to display when linking an Excel file to a parts file?
    - how *can* *pp* *get* *parameter* to *display* when *link* |Excel *file*| to |*part* *file*

Concept Understanding

- Trendelenburg>Trendelenberg>Airplane

- Glycated Hemoglobin > HA1C > Hemoglobin A1C > Diabetes > Diabetes Mellitus > Neuropathy > Retinopathy > Ketoacidosis

Results

- Unification of data across multiple platforms
- A virtual librarian making connections that might be missed
- Focus on authoritative content
- Extract greater value from enterprise data

Thank you!

- Questions?
- Paul Edelblut
  - pedelblut@email.vantage.com
  - +1-267-991-4435
  - @pauledelblut

NLP SUMMIT

HEALTHCARE

Presented by

John Snow LABS

www.nlpsummit.org