**“Colourless green ideas sleep furiously”: Approaching a new venture in NLP**

The principal objective of this module for me is to establish a foundation in research into formalising text into a data-structure to capture its meaning so that others can use what I’ve surveyed and deduced myself, then build upon it to ascend up this Mount Everest of an end goal.

The phrase ‘formalising text into a data-structure’ seems quite straightforward initially; shapes have been formalised like that thanks to research into Computer Vision so surely this can be repeated for text. However, Computer Vision and Graphics have been researched extensively; it is a mature field, whereas my research indicates that considerably less scholars have invested time into this aspect of Natural Language Understanding.

This blog describes my approaches, both successful and unsuccessful, to gathering information on a problem that Professor Teufel (University of Cambridge) describes as far from being solved; “The second wave of AI, right now, is soon going to fail because too much trickery and even self-trickery is used.”

All papers used in the process of making this blog are linked at the bottom.

**APPROACHES TO RESEARCH**

From the perspective of a first-year student such as myself, research papers are a nightmare. Between presumed understanding of advanced concepts in mathematics and descriptions of ideas that span several pages, distinguishing between valuable information and material of no relevance to me was always going to be challenging. In order to face that problem however, you need to acquire relevant research papers. This was the painfully arduous at times.

I started with keyword searching which could be considered a ‘lottery-based approach’ because while you are likely to find some relevant papers which include the terminology that you use, it’s possible that there are other insightful papers which are omitted from the search results because they address the same problem differently. I would actively discourage other researches from doing this for a long time because you’re unlikely to gather anything meaningful apart from the names of important people from the field of research. This in itself constitutes as something of value because it leads us directly to our next research approach: searching by people who have made a contribution to the field.

Keyword searching for phrases like ‘formalising text into data structure’ and ‘rule-based approach to formalising text’ yielded a limited supply of papers to work with. ‘Formalising natural language’ was surprisingly successful because I read a paper which lead me on to researching tools from Computational Linguistics. It described 4 tools for formalising grammars; XFST, GPSG, LFG and HPSG. Initially, I was hesitant to accept that researching linguistics papers would be of any value. However, I reminded myself that formalising text into a data structure is a problem which applies Machine Learning to a linguistics field of research.

I missed countless results that are related representing text by using the keywords selected above; they did not account for graph-based representations of the meaning of text which from my perspective, is a more semantically-oriented approach to

TLDR;

Use keyword searches and visit conference websites to identify those who have made valuable contributions to their field

**FIGURES OF INTEREST: NOAM CHOMSKY**

The first important name I came across while researching Linguistics is Avram Noam Chomsky (1928 – Present). In 1957, he published a monograph called ‘Syntactic Structures’ which argues that natural language semantics and syntax should be considered independent of each other; sentences which are syntactically coherent can have no meaning. His famous example is *‘Colourless green ideas sleep furiously’.* The syntactic structure [ADJ, ADJ, NOUN, VERB, ADV] is acceptable yet the individual words when put together are nonsense.

One valuable point that I deduced from reading about “Syntactic Structures” is that **a metric for success** for our working machine is for it to be able to **recognise nonsense**. The ability to do that would indicate to me that the machine is clever; it uses reasoning and understands each of the words in that sentence as well as why they make no sense in that order. A true measure of this intelligence would be if the machine could explain **why** that combination of words is invalid. In our example case, ‘Colourless green ideas sleep furiously’, a well-designed machine would argue [‘Something cannot be colourless and coloured at the same time.’, ‘Sleep is a passive state, so adverbs related to aggression don’t make sense’] just as a human would.

Furthermore, Chomsky distinguishes between a ‘grammatical’ and ‘ungrammatical’ sentences where ‘grammatical’ sentences are intuitively acceptable to a native speaker. His famous example is grammatical, but not statistically probable. He then relates phonology to semantics, arguing that how a word is said means more than the word itself.

**FIGURES OF INTEREST: CHRISTOPHER D. MANNING**

Professor Manning lectures and researches Natural Language Processing at the Stanford University. His focus is on a computational linguistics approach to parsing, Natural Language Inference and Multilingual Language Processing. In fact, he was a key developer in the SNLI corpus.

The paper describing the process of creating this corpus

*What is Natural Language Inference?* The task of determining if a hypothesis is true, undetermined or false given a premise. A metric for success for this is if a machine can determine the nature of the hypothesis having been shown only the premise and in addition, demonstrate ‘common sense’ by explaining why it has reached the decision.

**FIGURES OF INTEREST: SEBASTIAN PADÓ**

I found this researcher because he is a keynote speaker for RANLP which took place in Bulgaria during Summer 2019. His expertise lies in Computational Linguistics; having carried out postdoctoral research in Stanford. His research interests specifically address our ‘Everest’: developing representations for the meaning of natural language words, phrases and documents obtained from a corpora.

**WHY IS REPRESENTING THE MEANING OF WORDS IMPORTANT?**

The value of representing the meaning of words could be described anecdotally in terms of a band playing a song; your interpretation of a song is highest when all instruments are playing together. If you only listen to one instrument, your understanding of the overall song is greatly reduced. The same is true in NLU; if you isolate a word from its context, it is semantically ambiguous. Although context is key, syntactic ambiguity is another challenge that an intelligent machine would have to overcome. Take the Winograd Schema; a list of 150 sentences containing a referent (it) and accompanied by a couple of words, each changing what the referent refers to. If the machine can score highly on Winograd Schema examples, it’s a sign that it has common sense and contextual awareness.

**REQUIREMENTS FOR FORMALISING TEXT AS A DATASTRUCTURE 1: COMPUTATIONAL POWER**

In the paper ‘Language Modelling with Gated Convolutional Networks’, Dauphin, Fan, Auli and Grangier propose that a metric of success for a model. The throughput is the number of tokens being processed per unit time. Parallel processing accelerates this. Responsiveness is described as processing tokens in sequence. Typically, responsiveness and throughput are inversely proportional however batching can make them directly proportional to each other.

**INTERVIEW WITH DR. NILOOFER SHANAVAS (PwC OPERATE)**

1. Is your predoctoral academic background more centred around Computing or Linguistics?
2. Your research output includes tr-srw a term-weighting scheme designed to improve text classification performance.
3. Your paper explaining ‘tr-srw’
4. Are you aware of any alternative representations of the meaning behind text?