Natural Language Processing for Government: Problems and Potential

Preprint · April 2019

CITATIONS
2

READS
2,124

3 authors, including:

Nisansa de Silva
University of Moratuwa
78 PUBLICATIONS **515** CITATIONS
SEE PROFILE

Yashothara Shanmugarasa
UNSW Sydney
10 PUBLICATIONS **25** CITATIONS
SEE PROFILE

Natural Language Processing for Government: Problems and Potential

Yudhanjaya Wijeratne, Nisansa de Silva, Yashothara Shanmugarajah

LIRNE*asia* is a pro-poor, pro-market think tank whose mission is *Catalyzing policy change through research to improve people"s lives in the emerging Asia Pacific by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology.*

Contact: 12 Balcombe Place, Colombo 00800, Sri Lanka. +94 11 267 1160. info@lirneasia.net
www.lirneasia.net

This work was carried out with the aid of a grant from the International Development Research Centre (IDRC), Canada.

IDRC | CRDI
International Development Research Centre
Centre de recherches pour le développement international

Canada

1 Introduction

'It has not been for nothing that the word has remained man's principal toy and tool: without the meanings and values it sustains, all man's other tools would be worthless.'

- Lewis Mumford

Natural Language Processing (NLP) is a broad umbrella of technologies used for computationally studying large amounts of text and extracting meaning - both syntactic and semantic information. Software using NLP technologies, if engineered for that purpose, generally have the advantage of being able to process large amounts of text at rates greater than humans.

A large number of the functions of a government today revolve around vast amounts of text data - from interactions with citizens to examining archives to passing orders, acts, and bylaws. Under ideal conditions, NLP technologies can assist in the processing of these texts, thus potentially providing significant improvements in speed and efficiency to various departments of government.

Many proposals and examples exist illustrating how this can be done for multiple domains - from registering public complaints, to conversing with citizens, to tracking policy changes across bills and Acts.

This whitepaper seeks to examine both the current state of the art of NLP and illustrate government-oriented use cases that are feasible among resource-rich languages. Thus, in rough order:

1. A very brief examination of applications in government, using proposed use cases
2. The challenges of NLP and its present-day domain structure
3. The foundations required for these use cases to be implemented
4. Other problems, challenges and technical roadblocks to implementing these foundations that need to be solved before progress occurs

Before we begin, a disclaimer: this paper will focus on the more statistical aspects of NLP, as defined by the preamble that we will attempt to describe these various tendrils, but cannot claim fully representivity of everything possible. It is practically impossible to cover, in a short space, all possible roots and branches of the banyan tree that is NLP; we will instead base our discussion around uses that have been already proposed in the field of government.

Lastly, it should be noted that the vast majority of these use cases are designed for English text. Indeed, much of the NLP technology that exists is firmly centered around English and other languages from the Germanic-Romance family tree, owing to both a wealth of low-level language resources and researchers who gravitate towards the lingua franca. This has created a rift, with languages being divided into resource-rich and resource-poor languages [33, 53, 110, 122, 150]. Governments that need to work with resource-poor languages by choice or by necessity are thus at a relative disadvantage.

2 Thesis: a case for NLP

> *'There are only two things to learn in any language: the meaning of the words and the grammar. As for the meaning of the words, your man does not promise anything extraordinary; because in his fourth proposition he says that the language is to be translated with a dictionary. Any linguist can do as much in all common languages without his aid. I am sure that if you gave M. Hardy a good dictionary of Chinese or any other language, and a book in the same language, he would guarantee to work out its meaning.'*
>
> - Descartes to Mersenne, 20 November 1629 [37]

2.1 The State of the Art

The history of NLP is convoluted - it is commonly traced to Alan Turing's 1950 article "Computing machinery and intelligence", in which he proposed a machine that can understand human conversation [178]. Alternatively, it can be traced to the 1954 IBM-Georgetown experiment, which was a stab at the fully automated translation of Russian sentences into English [85, 115]. Both these efforts eventually spun themselves into the complex banyan tree that is NLP - a discipline that has tendrils in everything from Mersenne and Descartes' search for universal language structures [37] to Chomsky's theories of how existing languages can be reduced to rulesets [27].

Thankfully, its importance is easier to abstract if one approaches the field from the present instead of the past. Today, NLP is the quest to computationally understand meaning - both syntactic and semantic - from text.

For a piece of text to be understood, NLP systems must extract meaning at different levels, as shown by Liddy in her seminal 2001 book *Natural Language Processing* [97].

- **Phonological analysis:** The interpretation of language sounds. Liddy subdivides this into three types of rulesets: phonetic rules that capture sounds within words; phomenic rules that capture variations of pronunciation when words are chained together; prosidic rules that capture variations of stress and information across a sentence.
- **Morphological analysis**: The analysis of words at the smallest units of meaning, aka morphemes. In English, for example, this involves separating prefixes, suffixes and tense attachments to understand the information contained within.
- **Lexical analysis**: The analysis at the level of individual words - including the identification of parts of speech.
- **Syntactic analysis**: Analysis at the level of a sentence, parsing grammar structures to extract meaning. A key here is in understanding the subject and object of a sentence, and upon what an action is inflicted.

- **Semantic analysis**: The determination of the overall meaning of a sentence through the interactions between the word-level meanings identified previously. A key here is identifying the context in which a word is deployed in a sentence in a way that changes the overall meaning. Ie: running a program is different to running a marathon.

- **Discourse analysis**: The extraction of meaning from text units longer than a sentence. This also deals with determining the structure of a document and the function of a particular sentence - ie. whether it serves as the beginning, the middle or the end.

- **Pragmatic analysis**: The examination of context that is read into the contents of a text without being explicitly mentioned. Liddy explains this using two examples of the word "they" in sentences where the "they" reference different agents in the sentence. This requires pragmatic, or world knowledge, to solve.

Today's NLP breaks down various aspects of these tasks into sub-fields, each with their own extensive literature and approaches [149, 193]:

- **Information Retrieval (IR)** is the science of searching for and retrieving information within a document, corpus or snippet of text. The IR level is relatively rudimentary in that it usually simply matches text strings against queries - search engines [167] and a word processor's Ctrl+F function are good examples of this.

- **Information Extraction (IE)** is the domain of automatically extracting structured information from machine-readable documents. While this nominally sounds identical to IR, IE concerns itself with the transformation of unstructured or semi-structured data in documents into structured information, often making inferences with regard to domain and context. IE often uses techniques developed in IR as a base on which to perform this higher-order extraction of meaning.
  Notable subfields: Relationship extraction [109, 155, 156], Topic modelling [23, 183, 184], Sentiment analysis [54, 129, 163], Automatic summarization [61, 107, 124, 125, 182]

- **Natural Language Understanding (NLU)** concerns itself with computationally replicating a human-like understanding of language, often (but not unnecessarily) by reducing human language to a structured ontology. Language based reasoning systems are the general objective of these implementations. It falls into the infamous category of "AI-hard" problems. This domain often intergrates techniques from both the IR and IE fields.
  Notable subfields: Machine translation [16, 25, 130], discourse analysis [26, 108, 135, 144, 145, 195]

Liddy's text, while not built with the aforementioned classification in mind, rings remarkably prescient even in this reshaped context. She notes that "*Current NLP systems tend to implement modules to accomplish mainly the lower levels of processing. This is for several reasons. First, the application may not require interpretation at the higher levels. Secondly, the lower levels have been more thoroughly researched and implemented. Thirdly, the lower levels deal with smaller units of analysis, e.g. morphemes, words, and sentences, which are rule-governed, versus the higher levels of language processing which deal with texts and world knowledge, and which are only regularity-governed*" [97].

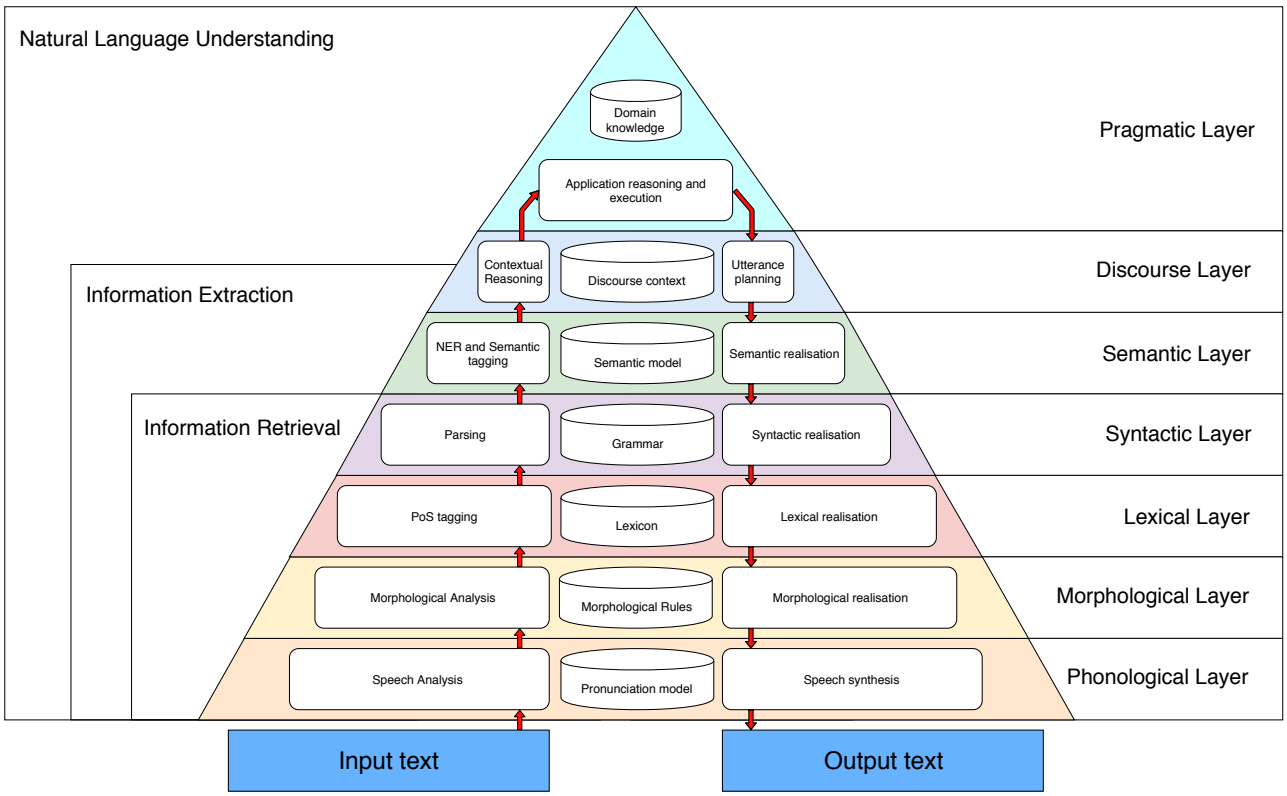

Figure 2.1: NLP layers and tasks

A graphical representation of these sections are shown in Figure 2.1. Note that the *discourse layer* straddles information extraction and natural language understanding.

This status quo, we feel, holds true as of the time of writing. Higher-level understanding is rare and still more efficiently performed by humans. NLP, despite over sixty years of research, is still an extraordinarily difficult task.

To appreciate this difficulty, consider this poem:

If the plural of man is always called men,
Why shouldn't the plural of pan be called pen?
If I speak of my foot and show you my feet,
And I give you a boot, would a pair be called beet?
If one is a tooth and a whole set are teeth,
Why shouldn't the plural of booth be called beeth?
In what other language do people recite at a play and play at a recital?
We ship by truck but send cargo by ship.
We have noses that run and feet that smell.
We park in a driveway and drive in a parkway.
And how can a slim chance and a fat chance be the same,
while a wise man and a wise guy are opposites?
You have to marvel at the unique lunacy of a language
in which your house can burn up as it burns down,
in which you fill in a form by filling it out, and
in which an alarm goes off by going on.

This is English, arguably the lingua franca of the world. It is a relatively simple language, efficient, and morphologically poorer than many Asian languages. And yet, as seen from this

simple poem above, it clearly has its own byzantine set of language rules, which somehow we humans manage to navigate despite their apparent muddled nature.

One approach to computationally extracting meaning from the poem would require the codification of intricate rulesets for every possible sub-domain. Indeed, initial natural language processing - powered in part by Chomskyan theory [27]- did indeed rely on such complex handwritten rules. However, as Jones and Nadkarni note, there are problems with this approach "*the rules may now become unmanageably numerous, often interacting unpredictably, with more frequent ambiguous parses (multiple interpretations of a word sequence are possible).*"

To demonstrate the issue, consider the following. The first two lines of the poem touches a very common musing on the English language; this is the question on the plurals of "*man*" versus "*human*". Man becomes "*men*" but human becomes "*humans*" instead of "*humen*". Now the reason for this is the fact that the "*man*" comes from the proto-Germanic roots and human comes from Latin roots. Human comes from Latin "*humanus*" which comes from *humi-anus* and further *homo-anus* [168]. *Homo-anus* has roots in *(dh)ghomon-anus* which goes back to *dhghem-anus* [127]. So, this means "*the one who belongs to those of earth*". This again is a dual meaning given that it can either be taken as "*as opposed to gods in sky*" or with a re-read of the myth of Prometheus making the first man in clay. On the other-hand "*man*" of today is short form of Germanic "*werman*" which means "*wer*" (man) "*man*" (person). This might be somewhat confusing to the speakers of modern English. This confusion can be solved by pointing out that English has one place "*wer*" survives with the meaning of "*man*". That is in "*werewolf*" which means "*man-wolf*". Alternatively, the origin of the word "*woman*" which comes from "*wifman*" can be analyzed. Which is more or less *wif* (woman) *man* (person) [169]. Hence, seemingly similar words in English have different rules on them depending on the origin. Thus the word "*man*" uses Germanic rule set to obtain the plural while "*human*" uses the Latin rules. A fully rule-based system will thus have to contain rules that derive the above etymological conclusion or brute-forced rules that dictate all forms and combinations of all words and parts.

Thus the various sub-fields of NLP have shifted from such limited rule-based approaches towards using massive language data and machine-learned models for various tasks. These data-based approaches are much less limited in what they can produce. Let us return to that English poem:

Figure 2.2: Part-of-Speech

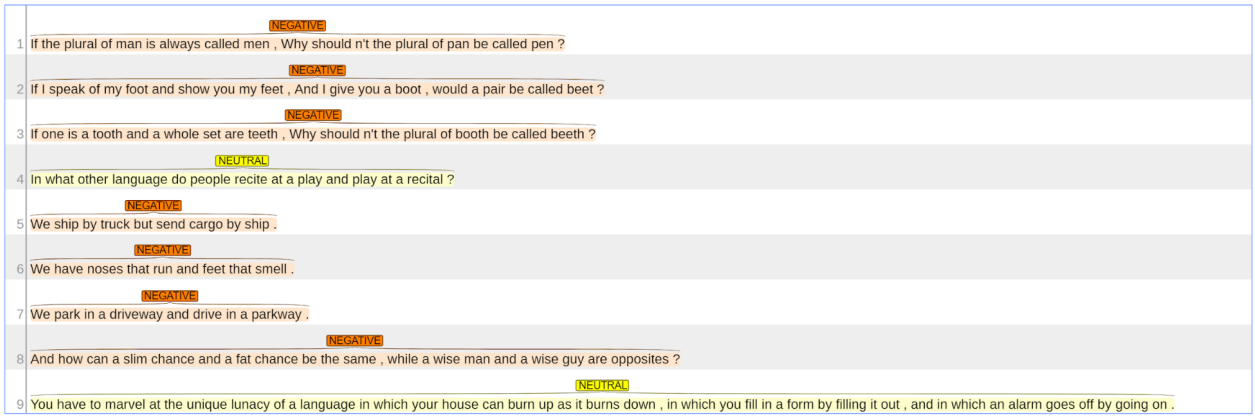

Figure 2.3: Sentiment Analysis

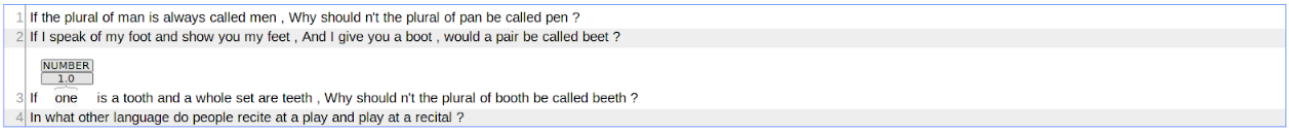

Figure 2.4: Named Entity Recognition

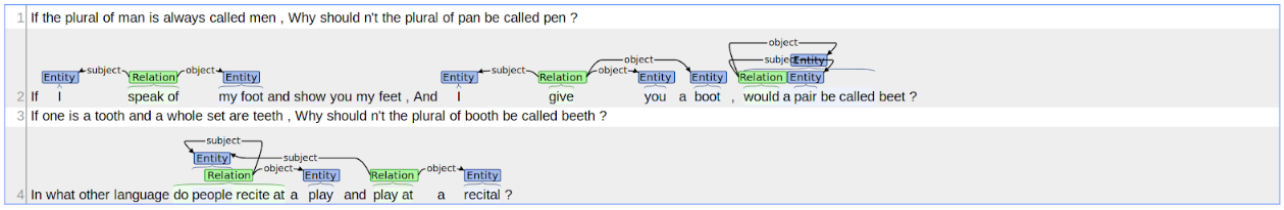

Figure 2.5: Open Information Extraction

Here, using a single, publicly available library; Stanford CoreNLP [109], we have attempted to examine the poem on various levels - lexical, syntactic, semantic. While not perfectly accurate, it has managed a reasonable analysis of the poem, which we know relies on multiple levels of wordplay. Under ideal conditions, NLP technologies such as this display extremely high rates of accuracy ($> 90\%$ in many cited instances) in domain-specific tasks. They also possess the ability to process large volumes of text data at great speed - the same program used for this one poem could be deployed on several hundred thousand poems, and it would return answers in a fraction of the time it would take a human to process the same.

These technologies are nearly ubiquitous in today's world. Search engines such as Google use NLP to process some 3.5 billion queries a day [4] to understand what the human on the other end is looking for. Smart Assistants such as Siri live in our phones and use such technologies to present information and carry out rudimentary conversations with users. Marketing and ad platforms use keyword analysis and word-matching to deliver advertisements at speed.

These applications have most prominently been the domain of private corporations and universities, but there exists a powerful argument for incorporating NLP into functions of government. Government functions revolve around vast amounts of text - from communiques to bills to Acts and suchlike. The analysis of text is an implicit aspect of a government official's or department's job.

While still incapable of truly human-level, higher-order analysis, NLP technologies are, under ideal conditions, able to both mimic and scale basic lower-order tasks that otherwise take up a great deal of human time and effort. With the right awareness of the pitfalls, and with human operators as a final check, NLP allows for a government that operates with more efficiency and speed than a purely human operation.

2.2 Applications

2.2.1 Electronic petitions, citizen proposals and complaints

Hagen et al. [60] of the University of Albany, New York, have studied petitions and federal policy suggestions from the *We the People* system launched by the White House under the Obama administration. They were able to extract topics from these petitions that held up well to qualitative analysis of those texts by humans; that is to say, the software inferred roughly the same subjects. They went on to expound the great use these techniques would have in the analysis of vast archives of public information - proposals, queries, complaints included. As they state, this allow governments to process "*large quantities of citizen generated policy suggestions through a largely automated process, with potential application to research on e-participation and policy informatics*." Furthermore, they were able to extract an informativeness score that may potentially be of use in identifying the most salient proposals in a topic; another metric, of course, being the number of signatures.

Similar work has been performed by Kowalski et al. [89], who used 145,000 reviews of over 7,000 care centers run by England's National Health Service (NHS) to demonstrate that NLP can be used as a very relevant mechanism for intergreting public feedback into government services. The UK government has, in fact, announced that they are using NLP for performing this very same function [63].

This kind of topic extraction allows not only an understanding of what large numbers of citizens care about, but also in sorting the relevant suggestions to the relevant department. Tjandra et al. [176] have proposed using a fairly well-known text classification algorithm that can take a complaint - in text, in the Indonesian language - and determine which government department it should be sent to based on the contents of the complaint.

2.2.2 Enhancing public digital archives and resource discovery interfaces

Another interesting use case comes from the UK Government's *Finding Things team* [3], which tackled the problem of creating a taxonomy for pages about education by using NLP techniques. They found that they could use topic modelling to substantially automate the process of assigning topics to documents - which is key for information search/discovery across digital archives. On many fronts, this process is similar to the analysis of petitions discussed above, except the results of the analysis contribute to some public-facing interface.

2.2.3 Automated crime reporting and government chatbots

Iriberri and Leroy [73], pointing out the need for crime reporting to be available 24/7, proposed an online crime reporting system which can extract relevant information from witnesses' narratives using natural language processing and ask additional questions using investigative interviewing techniques. This system ultimately feeds all the information thus acquired into a database for easier, more structured retrieval of information.

Systems like these open the door to more powerful automated crime reporting in general, a feature especially useful for countries and situations where police resources are under considerable strain. It is not infeasible to pair the ability to ask relevant questions over text with text-to-speech conversion, thus achieving the ability to automatically interview a witness over a voice channel.

However, even broader applications exist. This ability to extract useful information from a conversation, formulate questions that probe deeper into the issue, and then record said data into an appropriate mechanism can be applied to practically every citizen-facing aspect of government.

Singapore uses a bot[1] based on the Facebook Messenger platform which allows citizens to contact civil servants, find government-related news, and report issues to appropriate departments. A similar example is OpenDataKC[2], a government of Kansas City chatbot that can find information from Kansas's City's open data archive[3] on request. Travelbot[4], from *Transport for London* performs a similar Facebook-based service with regard to bus and train arrivals, route maps and line statuses.

A more complex example comes from the US Department of Homeland Security, which operates EMMA[5], a virtual assistant which handles requests regarding immigration services, green cards, and passports in both Spanish and English (including speech support). The City of Los Angeles, meanwhile, has built an application that allows people with Alexa devices to update themselves on news and events happening in city bounds [2]. The Dubai Electricity and Water authority operates RAMMAS[6], a multi-platform assistant which uses Google's AI platform.

As can be observed from these examples, much of the technology required for the creation of such NLP-powered interfaces is already available from technology giants that already have the necessary research and infrastructure required.

2.2.4 Examining government policy and news media

O'Halloran et al. [126] have proposed the use of NLP techniques for the study of financial regulation policy. Their research centers around scalable methods for classifying policy text by the information contained within, essentially automating a key requirement of traditional

[1] https://www.facebook.com/gov.sg/
[2] https://www.facebook.com/OpenDataKC/
[3] https://data.kcmo.org
[4] https://www.facebook.com/tfltravelbot/
[5] https://www.uscis.gov/emma
[6] https://www.dewa.gov.ae/en/rammas

studies: the need for annotators putting in thousands of person-hours into the extraction of information. Li et al. [94] have proposed using NLP in a similar fashion, extracting metrics that allow them to track to progress of policy ideas in bills related to the 2008 financial crisis. By identifying the appearance of topics in text, and then examining subsequent presentations of four bills, they were able to track trace the dominance of ideas - and identify which bills mutated the most, as well as policy ideas that were dropped as the bills progressed.

They later expanded their approach to model the evolution of the United States Code from 1926 to 2014 [95], identifying, '*the first appearance and spread of important terms in the U.S. Code like "whistleblower" and "privacy."*'

Similar topic modelling uses have been deployed to great effect on news. The BBC [1], not to mention multiple independent researchers [43, 93, 157] have demonstrated the use of topic models in news.

Nay [123] has proposed using NLP techniques in a novel fashion: Gov2Vec, which examines the legal corpus of an institution, and the vocabulary shared across institutions. Gov2Vec is thus able to discern differences between different government bodies. The method has also been extended to Presidential actions, Supreme Court decisions and summaries of Congressional bills. Perhaps the most useful use case for this technology would be as a research tool for discerning the minutae of large volumes of court cases, bills and acts and Parliamentary of Congressional sessions.

A project on ontology-based information extraction for the legal domain has worked on: instance population of an ontology using a legal case corpus through word embeddings [77, 78], deriving representative vectors for the said ontology [76], creating a domain specific semantic similarity measure for comparing legal documents [166], a legal document retrieval algorithm based on the above legal domain specific semantic similarity [167], an algorithm to identify relationships among sentences in court case transcripts using discourse relations [145], and a sentiment annotator for the legal domain with the objective of discovering statement bias [54]. All this have been done on a corpus made out of thousands of legal cases from the United States court system.

While all these use cases are oriented towards analysis in hindsight, this is fundamentally a task of information extraction and classification, which has already reared its head in previous use cases.

2.2.5 Capturing traffic and weather data from social media

If a particular government oversees a population that is sufficiently well-represented on social media, the detection of disruptive events becomes a possibility. In Japan, Sakaki et al. [151] of University of Tokyo have devised a system that can collect tweets and detect earthquakes with remarkable accuracy and email users - they cite that it picked up 96% of earthquakes (at a seismic intensity of 3 or above) than the Japan Meteorological Agency detected. Notably, their system worked faster than said agency when it came to notifying people of the threat.

This may be a reflection on the efficacy of meteorological agencies, or of the nature of the Japanese Twitterverse, but Japan is not the only country in which this may work. Teodorescu [173], a Romanian researcher, has compiled a much wider-ranging list of data sources which

might be used for detecting disasters, as well as a keyword based "*controlled grammar*" approach to filtering out such signals from the noise. Landwehr and Carley [92] of Carnegie Mellon go into extensive detail about what can be done with existing research, exploring case studies from Haiti to Uganda in an effort to paint a more comprehensive picture of the work done in this space. On the traffic front, a group of Sri Lankan researchers Athuraliya et al. [15] have examined crowdsourced traffic alerts (again, scraped out of Twitter). They take the analysis of data one step further by proposing (and implementing) an architecture that can capture and transform this raw text data from user into a machine-readable format - whereupon it can be entered in databases, or collected as statistics for road planners or police traffic departments.

What these point to is that many NLP technologies can, for a sufficiently verbose nation, be readily adapted scrape sites that host user-generated content. These systems, if set to look for the right signals, can work alongside existing systems to enforce better disaster response.

2.2.6 Public opinion and news monitoring

Lastly, Li [96], Wu and Chen [196] have explored the use of NLP technologies to monitor public opinion. Others have long since used topic modelling techniques, such as those discussed in previous use cases, to examine the spread of topics in news [1, 42].

While the context of these papers may be alarming to democratic institutions - one read "*the system presented in this paper can help government to correctly monitor the sensitive public opinion and guide them.*" - the applications remain; how citizens evaluate performance of public services may be fundamentally different from what organizational experts and decision-makers would understand [153]. The usefulness of using public opinion to better steer government service delivery cannot be overstated, though we do not encourage using such systems to "correct" citizens.

2.3 The Sum of All Parts

These use cases, considered as a whole, add strong technical evidence towards the feasibility of algorithmic governance in general [46]. While some "future government" technologies proposed by futurists are largely untested, require costly infrastructure overhauls, or easily supplanted by viable technologies, NLP represents a solid technical starting point which is both small enough to implement in micro-doses and yet powerful enough to make a significant impact [19].

It should be noted that the use cases discussed above are the ones that explicitly focus on some aspect of government, either by using data relevant to some mechanism of government, or by proposing a use case wherein these solutions can be used in a government department. In reality, ample NLP research exists that is widely cited, and tested on large datasets, often on news articles, academic research papers, or genetic data. Many of these can be quite easily adapted to government work simply by changing the data fed into them. Letting machines do the heavy lifting thus has historically allowed the human race to do more, faster.

3 Antithesis: the case against NLP

> *'But the Hebrew word, the word timshel—'Thou mayest'— that gives a choice. It might be the most important word in the world. That says the way is open. That throws it right back on a man. For if 'Thou mayest'—it is also true that 'Thou mayest not.'*
>
> - John Steinbeck, East of Eden

It may appear, at first glance, that all the research is in place, the technologies established, and that we are but a few clicks away from the next generation of government - happily powered by NLP, king of kings. But before we leap to implement, it behooves us to first understand this Ozymandias.

3.1 Context

The first, and most commonplace argument, is that NLP systems find it difficult to capture context. To illustrate, consider this sentence by Terry Pratchett, the author of Discworld: "*Humans need fantasy to be human. To be the place where the falling angel meets the rising ape*". This sentence packs an extraordinary amount of context, and, through it, subtlety. The reader would have to know that the "*falling angel*" references the Biblical Lucifer, falling out of favor with the Abrahamic God; they would also have to know that the "*rising ape*" is a reference to Charles Darwin's seminal book "*The Descent of Man, and Selection in Relation to Sex*" [31], and to the evolution of Homo Sapiens Sapiens; they would have to understand that when taken as a whole, these two sentences describe the historical conflict between Christianity and science, between mythology and rationality, and even - to some readers - of theism vs atheism. As of the time of writing, we must concur that no NLP system appears to be able to deal with such a sentence on this level of meaning. The current state of the art would miss the intended meaning of the sentence altogether.

3.2 Cooperativeness

The second argument concerns the nature of human communication, or rather, miscommunication. We refer here to Grice's Cooperative Principle [58]. Grice gives us four maxims for effective conversation:

- **The maxim of quantity**, where one tries to be as informative as one possibly can, and gives as much information as is needed, and no more.
- **The maxim of quality**, where one tries to be truthful, and does not give information that is false or that is not supported by evidence.

- **The maxim of relation**, where one tries to be relevant, and says things that are pertinent to the discussion.
- **The maxim of manner**, when one tries to be as clear, as brief, and as orderly as one can in what one says, and where one avoids obscurity and ambiguity.

Much miscommunication stems from violations of these principles, especially from obscurity and ambiguity, and human communication is rife with such; indeed, most of history can be boiled down to human miscommunciation. Nowhere is this tendency captured as well as in religion, where interpretations of the same text corpus lead many actors to split off and form entirely different movements. But lest this be interpreted as an issue with religion, we present this passage from a book by Richard Feyman, the legendary physicist [52]:

> "*A few years after I gave some lectures for the freshmen at Caltech (which were published as the Feynman Lectures on Physics), I received a long letter from a feminist group. I was accused of being anti-women because of two stories: the first was a discussion of the subtleties of velocity, and involved a woman driver being stopped by a cop. There's a discussion about how fast she was going, and I had her raise valid objections to the cop's definitions of velocity. The letter said I was making the women look stupid. The other story they objected to was told by the great astronomer Arthur Eddington, who had just figured out that the stars get their power from burning hydrogen in a nuclear reaction producing helium. He recounted how, on the night after his discovery, he was sitting on a bench with his girlfriend. She said, 'Look how pretty the stars shine!' To which he replied, 'Yes, and right now, I'm the only man in the world who knows how they shine.' He was describing a kind of wonderful loneliness you have when you make a discovery. The letter claimed that I was saying a women is incapable of understanding nuclear reactions.*"
>
> - Richard Feynman

The above conversation may be construed as a violation of Grice's Maxims: either Feynman violated the maxim of manner, or the letter-writers violated the maxim of quality, or both. Such violations happen extraordinarily often in human communication. As long as these maxims can be enforced on a conversation, NLP systems can be expected to do fairly well; however, the question whether they can be enforced or not. Grice himself points out that these maxims are not always easy to adhere to, and that upholding one may violate the other.

An NLP system interrogating witnesses, for example, would ideally be able to expect that they be truthful, and present its inputs (forms, text boxes etc) in such a way that the information is both relevant and of high quality.

Thus this offloads some of the burden onto the context of the system, the interfaces and formats in which language data is presented and recorded, and to the various incentives presented for information ordered in such a manner. Presumably the witnesses are aware that false testimony is punishable by law, and, driven by a need to return to their daily lives, they will attempt to be succinct in their statements instead of rambling on. Whether such incentives can be designed for every system in practice is a question that must be considered before deploying NLP systems for government purposes.

3.3 The capitalism of languages

The third argument is based on how different languages are, and this we would like to examine in a little more detail because of its direct implications on the majority of governments on this planet.

Languages differ greatly from each other in their their sentence structure (syntax), word structure (morphology), sound structure (phonology) and vocabulary (lexicon). These differences allow us to classify languages into families: in fact, one may visualize languages as a tree, as the artist Minna Sundberg did:

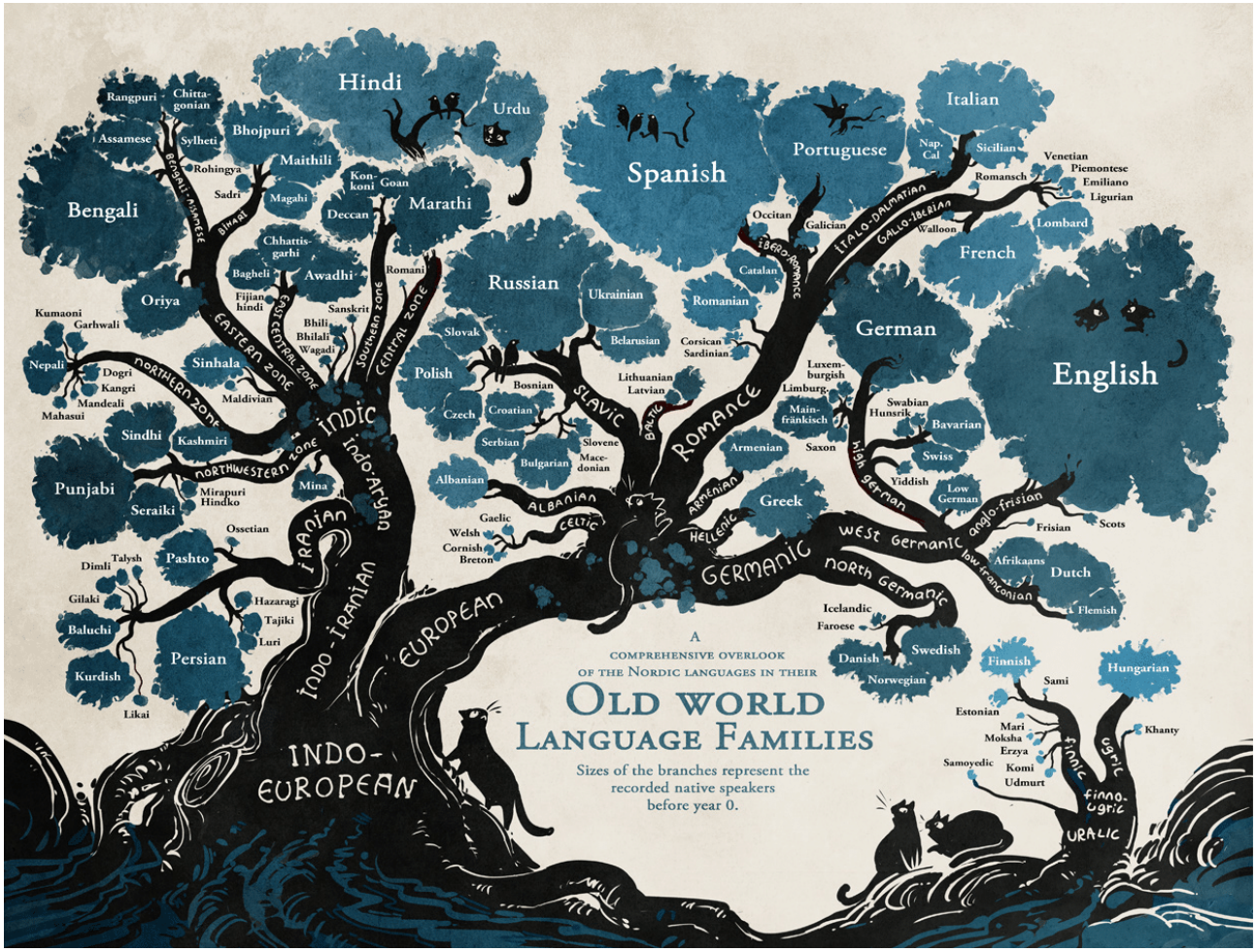

Figure 3.1: Indo-European Family of Languages [198]

Languages on the same branch resemble each other. As they diverge, the differences compound. For example, English, which belongs on the West Germanic tree, has three tenses - past, present and future. Sinhala, which belongs on the Indo-Aryan branch and is influenced heavily by Pali and Sankrit. Sinhala has only two tenses: the concept of past and not-past (*atheetha* and *anatheetha*).

Even on apparently close branches there exist major differences. Researchers working on a lexical ontological database (i.e. a WordNet) for Sinhala [192] examined Hindi, in the basis of their mutual closeness to Sanskrit found that the scripts were entirely different and there were very few words that could be translated between languages based on how similar they sounded. An algorithm written for one is unlikely to work in the other.

As a very visible example, imagine an algorithm that, having solved the intricacies of English, is then introduced to the Riau dialect of Sumatra. I quote from this article in the Atlantic[1], which apparently references research by David Gil [32]:

[1] https://www.theatlantic.com/international/archive/2016/06/complex-languages/489389/

"ayam" means chicken and "makan" means eat, but "Ayam makan" doesn't mean only "The chicken is eating." Depending on context, "Ayam makan" can mean the "chickens are eating", "a chicken is eating", "the chicken is eating", "the chicken will be eating", "the chicken eats", "the chicken has eaten", "someone is eating the chicken", "someone is eating for the chicken", "someone is eating with the chicken", "the chicken that is eating", "where the chicken is eating", and "when the chicken is eating".

These differences impact different NLP algorithms and sub-domains in different ways. Topic Modelling use cases, such as the one cited earlier, may likely pick up that there is something to do involving 'chicken' and 'eat'. However, Question Answering, such as those seen in the government chatbot use cases, are likely to be a much more difficult time with a language where "ayam makan", depending on the sentence context, may mean doing many different things involving a chicken.

There exists no NLP algorithms or use cases that can successfully navigate all these language differences. Attempts to build multilingual topic models [113] for even relatively similar languages show subtle differences in the topics highlighted- in the case of this cited paper, Danish, German, Greek, English, Spanish, Finnish, French, Italian, Dutch, Portuguese and Swedish translations of the *EuroParl* corpus[2] all expressed different results in the extraction of topics.

From this we may expect that algorithms already built will be able to work reasonably well within a language family, or a branch of the tree, but even so will display variations [7, 21]. This task will be far more difficult between branches - ranging from 'it's horribly inaccurate' to 'impossible'.

Which then brings us to the question: what languages are our NLP algorithms and use cases built for? The answer is overwhelmingly English [187].In the thesis section of this document, we mentioned examples that work "under ideal conditions". The ideal condition is simple: English. Of all the cases and papers cited above, the vast majority are performed with English text, using algorithms developed and tested on English text. The majority of these systems will not even function on ungrammatical English text, let alone a different language.

Today, the field of NLP has what practitioners call "resource-rich" and "resource-poor" languages [33, 53, 110, 122, 150]. Resource-poor languages are those that do not have the statistical resources required for ready analysis. Until the fundamental data is gathered, these are difficult nuts to crack. Resource-rich languages, on the other hand, are relatively low-hanging fruit and see more researchers working on them because much of the fundamental work is done - in much the same that more people started driving cars once the horseless carriage had been solved and made mainstream. Today, the vast amount of publicly available language data is overwhelmingly in English[3], which is recognized as the lingua franca of the sciences [44].

Thus we end up with a state of affairs where the vast majority of investment and dividends are focused on resource-rich languages and the resource-poor are still driving the language equivalent of horses on a superhighway. While both the need and use cases exist, the lack of data and research makes many languages impossible to analyze at the scales required [191].

This state of affairs would not be as much of problem if algorithms were language-agnostic. Unfortunately, the very nature of language tells us that they are not. To assume that all

[2] http://www.statmt.org/europarl/
[3] https://github.com/niderhoff/nlp-datasets

NLP algorithms designed for English will run perfectly, out-of-the-box, demands a state of advancement that we have not yet achieved.

3.4 A pessimistic summary

Therefore, the antithesis: despite the possibilities of NLP for governance, all of this is only optimal for, a handful of nations where the majority of the population is fluent in English, where both extensive research exists and practical application is possible. Not so much for the Phillipines, for whom Ethnologue lists 187 languages and dialects divided into 7 families, or India, whose constitution lists 22 languages as having official status. Ethnologue, of course, lists 448 living languages in India. The state of affairs is not as simple as Descartes envisioned it.

4 Synthesis

'For last year's words belong to last year's language And next year's words await another voice.'
- T.S. Eliot, Four Quartets

There are at least three ways to solve this state of affairs:

1. Map a path out of resource poverty and invest in re-creating fundamental NLP research for resource-poor languages - both data and algorithms
2. Work on machine translation models that can accurately render any language in English, thus unlocking the power of NLP
3. Design a universal proto-language (as proposed by Descartes) that does away with all the structural faults of English and can be the basis of all work in the future.

Various efforts are in place around the world explore the first two options. Universities around the world are sponsoring research to build the fundamental tools - tokenizers, lemmantizers etc - for low-level analysis. Preslav Nakov, at the Quatar Computing Research Institute, has conducted extensive work into adopting methods used for resource-rich languages into resource-poor variants. However, there are other factors at play here, and it serves us best to illustrate it via a review of literature from a microcosm (see Appendix A).

Meanwhile, Google, Facebook and Yandex are compiling vast corpuses of text data and training translation models between them. One should note that machine translation is still inaccurate even for these giants, especially for complex texts and structures that rely on wordplay (this should not be too surprising - it is difficult enough for humans versed in languages to accurately map Dante [74] and Borges [71] to English).

As a case in point, refer the Fig 4.1.

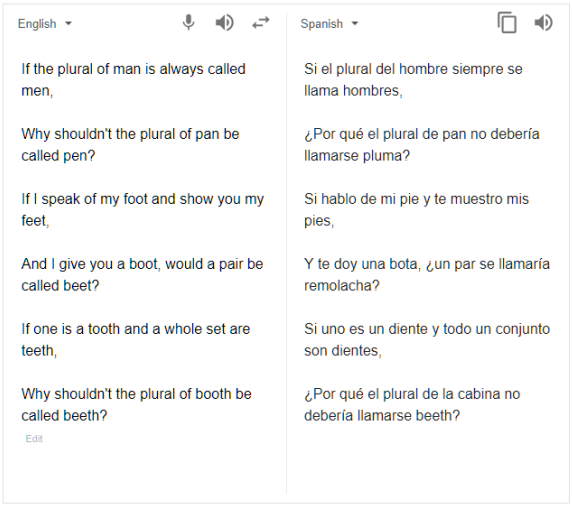

Figure 4.1: English to Spanish Translation

The second line in the Spanish translation actually reads "*Why should not the plural of **bread** be called a pen?*"

These difficulties, of course, are not relegated to English-Spanish translation. More complex pitfalls exist. Consider the word '*yuugen*'(幽玄) in Japanese . Yuugen appears to be a very complex concept that appears to have no direct translation to English. It may be to "*the beauty of gentle gracefulness*" [177] or to "*watch the sun sink behind a flower clad hill. To wander on in a huge forest without thought of return. To stand upon the shore and gaze after a boat that disappears behind distant islands. To contemplate the flight of wild geese seen and lost among the clouds. And, subtle shadows of bamboo on bamboo.*" It should be noted that both descriptions cite the same source - Zeami Motokiyo, one of the most prominent figures in Japanese Noh theatre. When one language contains concepts not found in the other, and are this difficult to define, translation is difficult.

Google, Facebook et al posit that despite these difficulties, there is an optimistic future now within our reach - with the right data, and supplementary tools, machine translation can be good enough that we can expect to unlock most of the power of NLP for any language, and that we may eventually figure out that Sumatran chicken or the Japanese playwright. This is essentially the realization of the goal set down by the 1954 IBM-Georgetown experiment.

While this push is convenient, it comes with caveats: governments may not necessarily want to rely sole on organizations for the enabling of these resources. Furthermore, such organizations may not have adequate language data to perform accurate analysis; the data and the knowledge thereof are likely to be more present in a region of native speakers. It is unrealistic to expect said corporations to hire native speakers in every language for this task, although Google and Facebook appear to be making remarkable headway through user-generated data.

As for the third option, it may very well be a pipe dream: while there exist artificial, theoretically universal languages - such as Solresol, Volapük, and Esperanto - making an entire populace adopt them appears to be a task beyond any government in place today. The Tower of Babel cannot be rebuilt.

4.1 Summary and recommendations

The ideal condition for NLP in governance, as of the time of writing, is to be a government that deals primarily in English, or to have a use case that relies only on English. Failing that, NLP implementations are currently best at resource-rich languages that fall within the Germanic and Romance families of languages, and can be rigorously benchmarked against the original English use case; this ensures that NLP use cases can be replicated with least modification.

Governments that deal primarily in resource-rich languages can readily implement these and thus have an advantage. Governments that deal with resource-poor languages are at a disadvantage. To quote William Gibson, "*The future is already here, it's just not evenly distributed*".

We suggest that the governments that can, do. As for the governments that cannot, we suggest partnerships with academia that can make continuous efforts towards developing the following:

1. **Publicly available text corpora in all national languages**: These will provide the basis of NLP research. Because of linguistic drift [154], these corpora will need to

be constantly evaluated and updated, and possibly classified into different structures in language where the formal and informal languages [8] differ considerably (i.e.: Sinhala).

2. **Tokenizers, Parts-of-speech taggers, lemmatizers, morphology analysers and parsers**: Tokenizers allow for the identification of individual words in a sentence; PoS-taggers allow the assignment of parts of speech to such words; lemmantizers allow the decomposition of words into their roots; parses allow the identification of noun and verb phrases. The rules that these constructs follow differ greatly from language to language, and thus these are best done with the aid of linguists (eg: in English, PoS-tagging is made more complicated by gerunds - "verbs ending in 'ing' that are used as nouns" [115]. In Sinhala and in Mandarin the equivalent structure is different).
3. **Machine-readable dictionaries**: These enable nouns and verbs to be translated into English for use in topic modelling and other applications that can sidestep complex language usage.

These efforts are unlikely to catapault the state of NLP into where it is in English overnight. However, if performed, they will enable NLP-assisted governance that does not wholly rely on large corporations that function outside a government's sphere of influence. Either way, we recommend that these initiatives be put in place as soon as possible, lest existing advantages and disadvantages compound. As shown in the thesis section, there is much to be gained from NLP; mastery of language, as Alexandre Dumas once said, does offer one the most remarkable opportunities.

Appendix A: A path out of resource poverty: pitfalls from Sinhala

Let us examine Sinhala language,a the native language of the largest ethnic group of Sri Lanka [20] [35]. Sinhala is spoken by a relatively small poluation and belongs same the Indo-European language tree as English [198], but in the Indo-Aryan branch.

Judging by research papers, it would seem that there are a significant number of implementations of NLP technologies for Sinhala, and that it should be ripe for use in the use cases outlined in the Thesis section. However, in the cases of almost all of the notable findings, the only thing that is publicly available for a researcher is a set of research papers. The corpora, tools, algorithm, and anything else that were discovered through these research are either locked away as properties of individual research groups, or worse, lost to time.

For any language, the key for NLP applications and implementations is the existence of adequate corpora. On this matter a relatively substantial Sinhala text corpus[1] was created by Upeksha et al. [180] by web crawling; a smaller Sinhala newes corpus[2] was created by de Silva [33].

WordNets [112] are extremely powerful and versatile components of many NLP applications. They encompass a number of linguistic properties that exist between the words in the lexicon of the language including but not limited to: hyponymy and hypernymy, synonymy, and meronymy. Their uses range from simple gazetteer listing applications [193] to information extraction based on semantic similarity [84, 197] or semantic oppositeness [34].

An attempt has been made to build a Sinhala Wordnet [192]. For a time it was hosted on [5], but this is now defunct and all the data and applications appear to be lost. However, even at its peak, due to the lack of volunteers for the crowd soured methodology of populating the WordNet, it was at best an incomplete product. Another effort to build a Sinhala Wordnet was initiated by Welgama et al. [189] independently from the above, but it too, appears to have ground to a halt.

As shown in Fig 2.1, morphological analysis is a ground level necessary component of natural language processing. Given that Sinhala is a highly inflected language [33, 88, 98], a proper morphological analysis process in vital.

However, the only work on this avenue of research which could be found was a study which was restricted to morphological analysis of Sinhala verbs [41]. There was no indication on whether this work was continued to cover other types of words. Completely independent of the above, Welgama et al. [190] attempted to evaluate machine learning approaches for Sinhala morphological analysis. Yet another independent attempt to create a morphological parser for Sinhala verbs was carried out by Fernando and Weerasinghe [49]. As a step on their efforts to create a system with the ability to do English to Sinhala machine translation, Hettige and

[1] https://osf.io/a5quv/
[2] https://osf.io/tdb84/

Karunananda [68] also claim to have created a morphological analyzer. Neither data nor tools from these have been made publicly available.

The next step after morphological analysis is a process known as Part of Speech (PoS) tagging. PoS tags differ in number and functionality from language to language. Therefore, the first step in creating an effective PoS tagger is to identifying the PoS tag set for the language.

In Sinhala, this work has been accomplished by Fernando et al. [51] and Dilshani et al. [40]. Expanding on that Fernando et al. [51] introduced a SVM Based PoS Tagger for Sinhala. Fernando and Ranathunga [50] provided an evaluation of different classifiers for the task of Sinhala PoS tagging. Several attempts to create a stochastic PoS tagger for Sinhala have been made, with work by Herath and Weerasinghe [62] and Jayasuriya and Weerasinghe [75] being most notable. A hybrid PoS tagger for Sinhala language was proposed by Gunasekara et al. [59].

Within another single group yet another set of studies was carried out to create a Sinhala PoS tagger, this time starting with the foundation of Jayaweera and Dias [79], which was then extended to a Hidden Markov Model (HMM) based approach [80] and an analysis of unknown words [81]. This group also presented a comparison of few Sinhala PoS taggers that are available to them [82].

While here it is obvious that there has been some follow up work after the initial foundation, it seems all of that has been internal to one research group at one institution as neither the data nor the tools of any of these findings have been made available for the use of external researchers.

Parsing is an area that is not completely solved, even in English, due to various ambiguities in natural language; however, in the case of English, there are systems that provide adequate results [109] if not perfect yet.

A prosodic phrasing model for sinhala language has been implemented by Bandara et al. [17]. While they do report reasonable results, yet again, do not provide any means for the public to access the data or the tools that they have developed. Work by Liyanage et al. [98] is also concentrated on this layer, given that they have worked on formalizing a computational grammar for Sinhala. Kanduboda and Prabath [88]'s work on Sinhala differential object markers also is an example of research done for the Sinhala language in the parser level. Another parser for the Sinhala language has been proposed by Hettige and Karunananda [69] with a model for grammar [67].

As shown in Fig 2.1, once the text is properly parsed, it can be processed using a Named-Entity-Recognition (NER) system. An NER system for Sinhala named *Ananya* has been developed by Manamini et al. [106].

But similar to the above developments, the developed data and tools seems to be held internally by the research group rather than making it publicly available. Another independent attempt on Sinhala NER has been done by Dahanayaka and Weerasinghe [29]; but that too is not accessible to the public.

A number of attempts have been made on semantic level applications for the Sinhala Language. A Sinhala semantic similarity measure has been developed for short sentences by Kadupitiya

et al. [86]. This work has been then extended by Kadupitiya et al. [87] for the use case of grading short answers. Data and tools for these projects are not publicly available.

Text classification is a popular application on the semantic layer of the NLP stack. Nanayakkara and Ranathunga [121] has implemented a system which uses corpus-based similarity measures for this propose. This too, is unavailable for external researchers.

A smaller implementation of Sinhala news classification has been attempted by de Silva [33]. As mentioned above, their news corpus is publicly available[3]; however, it is extremely small and thus may not provide much use for extensive research.

On the phonological layer, a Sinhala text-to-speech system was developed by Weerasinghe et al. [188]. However, it is not publicly accessible. A separate group has done work on Sinhala text to speech systems independently to the above [120].

On the converse, Nadungodage et al. [117] has work on Sinhala speech recognition with special notice given to Sinhala being a resource poor language. This project divides its focus on:continuity [116], active learning [118], and speaker adaptation [119].

A necessary component for the purpose of bridging Sinhala and English resources are English-Sinhala dictionaries.

The earliest and most extensive Sinhala-English dictionary available for consumption was by Malalasekera [104]. However, this dictionary is locked behind copyright and is not available for public research and development. The dictionary by Kulatunga [90] is publicly available for usage through an online web interface but does not provide API access or means to directly access the data set. The largest publicly available English-Sinhala dictionary data set is from a discontinued FireFox plug-in *EnSiTip* [162, 185] which bears a more than passing resemblance to the above. Hettige and Karunananda [64] claims to to have created a lexicon to help in their attempt to create a system capable of English to Sinhala machine translation.

Some work has been done by a group towards English to Sinhala translation as mentioned in some of the above paragraphs. This work includes; building a morphological analyzer [68], lexicon databases [64], a transliteration system [70], an evaluation model [66], a computational model of grammar [67], and a multi-agent solution [65]. Another group independently attempted English to Sinhala machine translation [99] with a statistical approach [100].

These studies highlight, in Sri Lanka, a fragmented approach among researchers that have caused these attempts not to cite or build upon the work of each-other. In many cases where similar work is done, it is a re-hashing on the same ideas adopted from resource rich languages - because of either the unavailability of data or lack of desire to refer and build on another group's work.

A study by de Silva [33], albeit in a limited context, shows a potential rapidly increasing the availability of resources via translation. In the study they attempted to use the resources available in a resource poor language (Sinhala) and then use simple word-by-word dictionary translation to obtain a crude English translation on which certain English language resources can be applied for the purpose of text classification. However, they reported that at the end this bridging was not needed given that Sinhala with simple NLP steps (up-to lexical layer

[3] https://osf.io/tdb84/

as shown in Fig 2.1) yielded better classification accuracy than English with advanced NLP steps (up-to semantic layer as shown in Fig 2.1), thus reducing the need to perform expensive bridging.

They further explain that this is due to the fact that Sinhala is a highly inflected language [88, 98] while English is a weakly inflected language [24]. We may infer that ideas which need higher order n-grams to be represented in English can easily be represented using unigrams or bigrams in Sinhala. A key lesson here is that some resource-poor languages, through intrinsic properties, may provide the developer with easier or more efficient approaches than a brute-force total conversion.

There is also potential to circumvent English by tapping into the much larger sphere of research in Tamil. Various translations have been attempted - there exist the government sponsored trilingual dictionary [36], and Weerasinghe and Dias [186] has created a multilingual place name database for Sri Lanka which may function both as a dictionary and a resource for certain NER tasks. Machine translation has also proved a boon: a neural machine translation for Sinhala and Tamil languages was initiated by Tennage et al. [170] [171, 172]. This project produced *Si-Ta* [142] a machine translation system of Sinhala and Tamil official documents. In the statistical machine translation front, Farhath et al. [48] worked on integrating bilingual lists. Attempts by Weerasinghe [187] and Sripirakas et al. [165] also focused on statistical machine translation while Jeyakaran [83] attempted a kernel regression method. Yet another attempt was made by Pushpananda et al. [133] which they later extended with some quality improvements [134].

However, attempts to utilize Sinhala and Tamil fall short of the mark of raising either or both languages to the level of resource rich languages. Again, the fault lie in the scarcity of tools and applications as well as the unavailability of the those few that actually exist. Thus, Sinhala persists in a state of resource poverty, and provides an abject lesson of practical pitfalls that can arise in NLP technology research.

Appendix B: Lessons from a language in development: Tamil

We present Tamil here as another case study of a resource-poor language, but one that is not limited to an island with a small population. Tamil is the native language of around 74 million people in Sri Lanka, Singapore, the Tamil Nadu state of India, and other small groups originated from those areas [6]. Unlike Sinhala, which belongs to the same Indo-European language tree as English [198], Tamil belongs to the Dravidian language tree [6]. Just like Sinhala, Tamil script is also a descendant of the Indian Brahmi script [30], and belongs to the Aramaic family of scripts [47, 152].

Tamil sees substantially larger corpora efforts in place than Sinhala. The Forum for Information Retrieval Evaluation (FIRE) [103], through the use of web crawling, has assembled a sizeable corpus of Tamil text. The Tamil Virtual Academy has created a a corpora from Sangam, Medieval and Modern Tamil - a corpus of some 150 Million words[1]. The Central Institute of Indian Languages, which also focused on creating corpora for Indian languages, has a corpus of 10.9 million works [2] contains 10,933,484 words. Furthermore, the Jawaharlal nehru University has, as of the time of writing, a of some 30,000 sentences. A Tamil monolingual corpus[3] has been created by University of Pennsylvania by collecting text from two short novels Ramasamy et al. [141], Goldhahn et al. [57] and IIIT-Hyderabad [179] are also, to the best of our knowledge, working on creating corpus for Tamil language. To the best of our knowledge, universities appear to be collaborating on the greater task of assembling corpuses for languages spoken in India.

Tamil dictionary efforts appear to be rather difficult to use for research or large-scale NLP. The earliest and most extensive Tamil-English dictionary available for consumption was by Winslow and Knight [194]. However, this dictionary is only available online in image format. There are online dictionaries such as Tamilcube[4], tamildict[5], Oxford Tamil[6], Agarathi[7], Glosbe[8], Kapruka[9] available for English-Tamil language, but most of these projects do not provide API access or some means of directly accessing the data.

A few attempts have been made to build a Tamil Wordnet. There exists an open source Wordnet on a website by Rajendran et al. [139]. The team behind this project is now working on Wordnet for Tamil, which aims to build a bigger Dravidian WordNet. [137]. Another effort to build a Tamil Wordnet was done by Anna University [28], attempting to capture the relationships between 50,000 root words in Tamil.

[1] http://www.tamilvu.org/en/tamil-corpus-bank/
[2] http://www.ldcil.org/resourcesTextCorp.aspx
[3] http://ccat.sas.upenn.edu/plc/tamilweb/
[4] http://dictionary.tamilcube.com/
[5] http://www.tamildict.com/english.php
[6] https://ta.oxforddictionaries.com/
[7] https://agarathi.com/
[8] https://en.glosbe.com/en/ta/
[9] https://www.kapruka.com/dictionary/EnglishToSinhala.jsp

Tamil has multiple lines of research into morphological analysis. Different technologies have been used in creating morphological analyzers: [22, 147] developed tools as far back as 2001, and the Resource Centre for Indian Language Technological Solutions (RCILTS) - Tamil has prepared a morphological analyzer ('Atcharam') [10]. Anand Kumar et al. [9], Duraipandi [45], Ganesan [55], Menon et al. [111], Parameshwari [131], Rajan et al. [136], Rao and Prameshwari [143] have individually taken different tracks and applied different technologies on Tamil corpuses, thereby providing a useful view of potential technologies and dead-ends in the field.

Recently two research works have focused on Sri Lankan Tamil. Lushanthan et al. [102] worked on building morphological analyzer by rewriting the lexicon and the orthographic rules of Tamil language as regular expressions. And, Tamil morphological analyzer using support vector machines was built by Mokanarangan et al. [114] which gives better result compare to [102] .

As with WordNets, various methodologies have yielded progress on the PoS front. A POS tagger based on a phonological approach was proposed by Renganathan [147]. Arulmozhi et al. [14] developed a POS tagger, albeit a limited one, for Tamil using a rule-based approach. A hybrid POS tagger for Tamil using Hidden Markov Models and a ruleset was developed by Arulmozhi and Sobha [13]. Ganesan [55], Selvam and Natarajan [159] and [38] also proposed and developed PoS taggers. Anna University developed a POS tagger [28] using a tagset standardized by the Government of India and Government of Tamil Nadu. Most recently, Mokanarangan et al. built a POS tagger using the Graph based Semi-Supervised approach [174]. This also highlighted a difference in dialects, giving better results for a Sri Lankan Tamil corpus compared to other taggers.

There are few attempts to build a parser for Tamil language. Early work by Rajendran [138] concentrated on this layer; Dhivya et al. [39] focused on clauses. Ramasamy and Žabokrtskỳ [140] used rule-based and corpus-based approaches that, in initial tests, achieved relatively high percentages of accuracy for the tasks tested on. Ariaratnam et al. [12] proposed a newer shallow parser. However, it should be noted that both the latter projects, which can be considered quite advanced, still yield accuracy rates lower than 75 percent.

Named Entity Recognition faces much difficulty, as it is both domain specific and changes depending on dialect.Domain-specific NER includes work by Vijayakrishna and Sobha [181] (for tourism), by Antony and Mahalakshmi [11] (for medical documents) and Theivendiram et al. [175] (for official documents in Sri Lankan Tamil). The majority of the research is centered on Indian Tamil dialects, such as work by Pandian et al. [128], Malarkodi et al. [105], Srinivasagan et al. [164].

A number of attempts have been made on semantic level applications for the Tamil Language, but they are not necessarily replicable. A document summarization for Tamil language has been developed using semantic graph method by Banu et al. [18]. The sentiment analysis field has yielded a lot of work in this regard: Patra et al. [132] focused on tweet data sentiment analysis, Ravishankar [146] worked on sentiment analysis on a corpus regarding Tamil moviesl Sharmista and Ramaswami [161] implemented a tree based opinion mining system in Tamil for product recommendations, and Se et al. [158] used machine learning to predict sentiment reviews for Tamil movies. Short Tamil sentence similarity calculation using knowledge-based and corpus-based similarity measures was developed by Selvarasa et al. [160]. However, all these datasets remain unavailable for use by external researchers.

Various works with different approaches have been proposed for translation between Tamil-

English languages. In general, they tend to work well for shorter and regular sentences, and lose accuracy as sentences become longer and more irregular.

Statistical Machine Translation (SMT) has, in particular, been popular for tackling translation Germann [56] and Renganathan [148] are some of the earliest sources using traditional SMT. Later, Loganathan [101] developed another SMT system by integrating morphological information. Kumar et al. [91] also used SMT by applying manually created reordering rules to syntactic trees, which led to a performance improvement.

The elephant in the room us, of course, Google's neural machine translation approach, which uses massive amounts of text data and works reasonable well for Tamil-English translation.

These studies highlight both the history and the status quo of research into Natural Language Processing in Tamil. While initially research happened in isolation, as in Sri Lanka, concerted efforts are now being made for collaboration between different universities [72] and through longstanding projects like Code Mix Entity Extraction in Indian Languages (CMEE-IL).

However, Tamil highlights two problems: one, is with dialects. While most most of the time we consider Tamil as one big umbrella, there are some important divergences between these two dialects of Tamil that crop up when dealing with Indian Tamil as opposed to Sri Lankan Tamil.

The other is that even with these efforts, there is a profound scarcity of practical applications, especially when compared to English. The majority of work remains theoretical, and in the realm of finding more accurate approaches, and thus is still decades behind the status quo in English, Thus, Tamil remains a low-resource language, although unlike Sinhala it appears to be on a more optimistic trajectory, and thus provides useful lessons for understanding the growth of technologies in resource-poor languages.

Bibliography

[1] Topic modeling in the news. URL http://bbcnewslabs.co.uk/projects/topic-modeling/.

[2] Introducing... alexa for the city of los angeles! | city of los angeles. URL https://www.lacity.org/blog/introducing-alexa-for-city-los-angeles.

[3] Using data science to build a taxonomy for gov.uk - data in government. URL https://dataingovernment.blog.gov.uk/2017/01/12/using-data-science-to-build-a-taxonomy-for-gov-uk/.

[4] Google search statistics. URL http://www.internetlivestats.com/google-search-statistics/.

[5] Sinhala wordnet. URL http://www.wordnet.lk/.

[6] Ethnologue. URL https://www.ethnologue.com/language/tam.

[7] Language families and nlp - chatbot pack. URL https://www.chatbotpack.com/language-families-nlp/.

[8] Formal and informal language. URL https://goo.gl/yuvd3k.

[9] M Anand Kumar, V Dhanalakshmi, KP Soman, and S Rajendran. A sequence labeling approach to morphological analyzer for tamil language. *IJCSE) International Journal on Computer Science and Engineering*, 2(06):1944–195, 2010.

[10] P Anandan, K Saravanan, Ranjani Parthasarathi, and TV Geetha. Morphological analyzer for tamil. In *International Conference on Natural language Processing*, 2002.

[11] J Betina Antony and GS Mahalakshmi. Named entity recognition for tamil biomedical documents. In *Circuit, Power and Computing Technologies (ICCPCT), 2014 International Conference on*, pages 1571–1577. IEEE, 2014.

[12] I Ariaratnam, AR Weerasinghe, and C Liyanage. A shallow parser for tamil. In *Advances in ICT for Emerging Regions (ICTer), 2014 International Conference on*, pages 197–203. IEEE, 2014.

[13] P Arulmozhi and L Sobha. A hybrid pos tagger for a relatively free word order language. In *Proceedings of the First National Symposium on Modeling and Shallow Parsing of Indian Languages*, pages 79–85, 2006.

[14] P Arulmozhi, L Sobha, and B Kumara Shanmugam. Parts of speech tagger for tamil. In *Symposium on Indian Morphology, Phonology & Language Engineering, March*, pages 19–21, 2004.

[15] CD Athuraliya, MKH Gunasekara, Srinath Perera, and Sriskandarajah Suhothayan. Real-time natural language processing for crowdsourced road traffic alerts. In *Advances in ICT for Emerging Regions (ICTer), 2015 Fifteenth International Conference on*, pages 58–62. IEEE, 2015.

[16] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[17] WMC Bandara, VMS Lakmal, TD Liyanagama, SV Bulathsinghala, Gihan Dias, and Sanalh Jayasena. A new prosodic phrasing model for sinhala language. 2013.

[18] M Banu, C Karthika, P Sudarmani, and TV Geetha. Tamil document summarization using semantic graph method. In / *iccima*, pages 128–134. IEEE, 2007.

[19] Medha Basu. How natural language processing will change governments | govinsider, 2018. URL https://govinsider.asia/inclusive-gov/natural-language-processing-explainer/.

[20] Laurie Bauer. *Linguistics Student's Handbook*. Edinburgh University Press, 2007.

[21] Emily M Bender. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6(3):1–26, 2011. URL https://pdfs.semanticscholar.org/105e/d573024e9a31eddc766b6018297ab4383bb9.pdf.

[22] Akshar Bharati, Rajeev Sangal, Sushma Bendre, Pavan Kumar, and KR Aishwarya. Unsupervised improvement of morphological analyzer for inflectionally rich languages. In *NLPRS*, pages 685–692, 2001.

[23] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[24] Laurel J Brinton. *The structure of modern English: A linguistic introduction*, volume 1. John Benjamins Publishing, 2000.

[25] Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

[26] Lynn Carlson, Mary Ellen Okurowski, and Daniel Marcu. *RST discourse treebank*. Linguistic Data Consortium, University of Pennsylvania, 2002.

[27] Noam Chomsky. *Aspects of the Theory of Syntax*, volume 11. MIT press, 2014.

[28] Computational Linguistics Research Group - AU-KBC Research Centre. Lexical resources. URL http://www.au-kbc.org/nlp/lex_re.html.

[29] JK Dahanayaka and AR Weerasinghe. Named entity recognition for sinhala language. In *Advances in ICT for Emerging Regions (ICTer), 2014 International Conference on*, pages 215–220. IEEE, 2014.

[30] Peter T Daniels and William Bright. *The world's writing systems*. Oxford University Press on Demand, 1996.

[31] Charles Darwin. *The descent of man and selection in relation to sex*, volume 1. Murray, 1888.

[32] GU David. Word order without syntactic categories how riau indonesian does it. *Verb first: On the syntax of verb-initial languages*, 73:243, 2005.

[33] Nisansa de Silva. Sinhala text classification: Observations from the perspective of a resource poor language. 2015.

[34] Nisansa de Silva, Dejing Dou, and Jingshan Huang. Discovering inconsistencies in pubmed abstracts through ontology-based information extraction. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 362–371. ACM, 2017.

[35] Department of Census and Statistics Sri Lanka. Percentage of population aged 10 years and over in major ethnic groups by district and ability to speak sinhala, tamil and english languages. URL https://goo.gl/nnVZSd.

[36] Department of Official Languages, Sri Lanka. Tri-lingual dictionary. URL https://www.trilingualdictionary.lk/.

[37] René Descartes and Anthony John Patrick Kenny. Philosophical letters. 1970.

[38] V Dhanalakshmi, G Shivapratap, and Rajendran S Soman Kp. Tamil pos tagging using linear programming. 2009.

[39] R Dhivya, V Dhanalakshmi, M Anand Kumar, and KP Soman. Clause boundary identification for tamil language using dependency parsing. In *International Joint Conference on Advances in Signal Processing and Information Technology*, pages 195–197. Springer, 2011.

[40] N Dilshani, S Fernando, S Ranathunga, S Jayasena, and G Dias. A comprehensive part of speech (pos) tag set for sinhala language. The Third International Conference on Linguistics in Sri Lanka, ICLSL 2017 ..., 2017.

[41] WSN Dilshani and G Dias. A corpus-based morphological analysis of sinhala verbs. The Third International Conference on Linguistics in Sri Lanka, ICLSL 2017 ..., 2017.

[42] Kevin Dooley, Steven Corman, and Dan Ballard. Centering resonance analysis: A superior data mining algorithm for textual data streams. Technical report, CRAWDAD TECHNOLOGIES LLC CHANDLER AZ, 2004.

[43] Kevin J Dooley and Steven R Corman. The dynamics of electronic media coverage. *Communication and terrorism: Public and media responses to*, 9(11):121–135, 2002. URL http://www.public.asu.edu/~corman/dist/dooley_corman_comm_terror.pdf.

[44] David G Drubin and Douglas R Kellogg. English as the universal language of science: opportunities and challenges, 2012.

[45] R Duraipandi. The mophological generator and parsing engines of tamil verb forms. *Tamil Internet*, 2006.

[46] Zeynep Engin and Philip Treleaven. Algorithmic government: Automating public services and supporting civil servants in using data science technologies. *The Computer Journal*, page bxy082, 2018. doi: 10.1093/comjnl/bxy082. URL http://dx.doi.org/10.1093/comjnl/bxy082.

[47] Harry Falk. *Schrift im alten Indien: ein Forschungsbericht mit Anmerkungen*, volume 56. Gunter Narr Verlag, 1993.

[48] Fathima Farhath, Surangika Ranathunga, Sanath Jayasena, and Gihan Dias. Integration of bilingual lists for domain-specific statistical machine translation for sinhala-tamil. In *2018 Moratuwa Engineering Research Conference (MERCon)*, pages 538–543. IEEE, 2018.

[49] Niroshinie Fernando and Ruwan Weerasinghe. A morphological parser for sinhala verbs. In *Proceedings of the International Conference on Advances in ICT for Emerging Regions*, 2013.

[50] Sandareka Fernando and Surangika Ranathunga. Evaluation of different classifiers for sinhala pos tagging. In *2018 Moratuwa Engineering Research Conference (MERCon)*, pages 96–101. IEEE, 2018.

[51] Sandareka Fernando, Surangika Ranathunga, Sanath Jayasena, and Gihan Dias. Comprehensive part-of-speech tag set and svm based pos tagger for sinhala. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 173–182, 2016.

[52] Richard Phillips Feynman and Ralph Leighton. *" What do you care what other people think?": further adventures of a curious character*. WW Norton & Company, 2001.

[53] Mohamed H. Gad-Elrab, Mohamed Amir Yosef, and Gerhard Weikum. Named entity disambiguation for resource-poor languages. In *Proceedings of the Eighth Workshop on Exploiting Semantic Annotations in Information Retrieval*, ESAIR '15, pages 29–34, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3790-8. doi: 10.1145/2810133.2810138. URL http://doi.acm.org/10.1145/2810133.2810138.

[54] Viraj Gamage, Menuka Warushavithana, Nisansa de Silva, Amal Shehan Perera, Gathika Ratnayaka, and Thejan Rupasinghe. Fast approach to build an automatic sentiment annotator for legal domain using transfer learning. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 260–265, 2018.

[55] M Ganesan. Morph and pos tagger for tamil. *Software) Annamalai University, Annamalai Nagar*, 2007.

[56] Ulrich Germann. Building a statistical machine translation system from scratch: how much bang for the buck can we expect? In *Proceedings of the workshop on Data-driven methods in machine translation-Volume 14*, pages 1–8. Association for Computational Linguistics, 2001.

[57] Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *LREC*, volume 29, pages 31–43, 2012.

[58] H Paul Grice. Logic and conversation. *1975*, pages 41–58, 1975.

[59] Dilmi Gunasekara, WV Welgama, and AR Weerasinghe. Hybrid part of speech tagger for sinhala language. In *Advances in ICT for Emerging Regions (ICTer), 2016 Sixteenth International Conference on*, pages 41–48. IEEE, 2016.

[60] Loni Hagen, Teresa M. Harrison, Özlem Uzuner, Tim Fake, Dan Lamanna, and Christopher Kotfila. Introducing textual analysis tools for policy informatics: A case study of e-petitions. In *Proceedings of the 16th Annual International Conference on Digital Government Research*, dg.o '15, pages 10–19, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3600-0. doi: 10.1145/2757401.2757421. URL http://doi.acm.org/10.1145/2757401.2757421.

[61] Udo Hahn and Inderjeet Mani. The challenges of automatic summarization. *Computer*, 33(11):29–36, 2000.

[62] Dulip Lakmal Herath and AR Weerasinghe. A stochastic part of speech tagger for sinhala. In *Proceedings of the 06th International Information Technology Conference*, pages 27–28, 2004.

[63] Dan Heron. Understanding more from user feedback - data in government. URL https://dataingovernment.blog.gov.uk/2016/11/09/understanding-more-from-user-feedback/.

[64] B Hettige and AS Karunananda. Developing lexicon databases for english to sinhala machine translation. In *Industrial and Information Systems, 2007. ICIIS 2007. International Conference on*, pages 215–220. IEEE, 2007.

[65] B Hettige, AS Karunananda, and G Rzevski. A multi-agent solution for managing complexity in english to sinhala machine translation. *Complex Systems: Fundamentals & Applications*, 90:251, 2016.

[66] Budditha Hettige and S Karunananda Asoka. An evaluation methodology for english to sinhala machine translation. In *Information and Automation for Sustainability (ICIAFs), 2010 5th International Conference on*, pages 31–36. IEEE, 2010.

[67] Budditha Hettige and AS Karunananda. Computational model of grammar for english to sinhala machine translation. In *Advances in ICT for Emerging Regions (ICTer), 2011 International Conference on*, pages 26–31. IEEE, 2011.

[68] Budditha Hettige and Asoka S Karunananda. A morphological analyzer to enable english to sinhala machine translation. In *Information and Automation, 2006. ICIA 2006. International Conference on*, pages 21–26. IEEE, 2006.

[69] Budditha Hettige and Asoka S Karunananda. A parser for sinhala language-first step towards english to sinhala machine translation. In *Industrial and Information Systems, First International Conference on*, pages 583–587. IEEE, 2006.

[70] Budditha Hettige and Asoka S Karunananda. Transliteration system for english to sinhala machine translation. In *Industrial and Information Systems, 2007. ICIIS 2007. International Conference on*, pages 209–214. IEEE, 2007.

[71] Andrew Hurley. What i lost when i translated jorge luis borges. URL https://core.ac.uk/download/pdf/25683993.pdf.

[72] Indian Institute of Technology Patna. Ai-nlp-ml group. URL http://www.iitp.ac.in/~ai-nlp-ml/.

[73] A. Iriberri and G. Leroy. Natural language processing and e-government: Extracting reusable crime report information. In *2007 IEEE International Conference on Information Reuse and Integration*, pages 221–226, Aug 2007. doi: 10.1109/IRI.2007.4296624.

[74] Clive James. Clive james on translating dante - telegraph. URL https://www.telegraph.co.uk/culture/books/booknews/10148152/Clive-James-on-translating-Dante.html.

[75] M Jayasuriya and AR Weerasinghe. Learning a stochastic part of speech tagger for sinhala. In *Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on*, pages 137–143. IEEE, 2013.

[76] Vindula Jayawardana, Dimuthu Lakmal, Nisansa de Silva, Amal Shehan Perera, Keet Sugathadasa, and Buddhi Ayesha. Deriving a representative vector for ontology classes with instance word vector embeddings. *arXiv preprint arXiv:1706.02909*, 2017.

[77] Vindula Jayawardana, Dimuthu Lakmal, Nisansa de Silva, Amal Shehan Perera, Keet Sugathadasa, Buddhi Ayesha, and Madhavi Perera. Semi-Supervised Instance Population of an Ontology using Word Vector Embeddings. In *Advances in ICT for Emerging Regions (ICTer), 2017 Seventeenth International Conference on*. IEEE, September 2017.

[78] Vindula Jayawardana, Dimuthu Lakmal, Nisansa de Silva, Amal Shehan Perera, Keet Sugathadasa, Buddhi Ayesha, and Madhavi Perera. Word vector embeddings and domain specific semantic based semi-supervised ontology instance population. *ICTer*, 11(1), 2018.

[79] AJPMP Jayaweera and NGJ Dias. Part of speech (pos) tagger for sinhala language. 2011.

[80] AJPMP Jayaweera and NGJ Dias. Hidden markov model based part of speech tagger for sinhala language. *arXiv preprint arXiv:1407.2989*, 2014.

[81] AJPMP Jayaweera and NGJ Dias. Unknown words analysis in pos tagging of sinhala language. In *Advances in ICT for Emerging Regions (ICTer), 2014 International Conference on*, pages 270–270. IEEE, 2014.

[82] Manoj Jayaweera and NGJ Dias. Comparison of part of speech taggers for sinhala language. 2016.

[83] Mahendran Jeyakaran. A novel kernel regression based machine translation system for sinhala-tamil translation. 2013.

[84] Jay J Jiang and David W Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc of 10th International Conference on Research in Computational Linguistics, ROCLING'97*. Citeseer, 1997.

[85] Karen Sparck Jones. Natural language processing: a historical review. In *Current issues in computational linguistics: in honour of Don Walker*, pages 3–16. Springer, 1994.

[86] JCS Kadupitiya, Surangika Ranathunga, and Gihan Dias. Sinhala short sentence similarity calculation using corpus-based and knowledge-based similarity measures. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WSSANLP2016)*, pages 44–53, 2016.

[87] JCS Kadupitiya, Surangika Ranathunga, and Gihan Dias. Sinhala short sentence similarity measures using corpus-based simi-larity for short answer grading. In *6th Workshop on South and Southeast Asian Natural Language Processing*, pages 44–53, 2017.

[88] A Kanduboda and B Prabath. On the usage of sinhalese differential object markers object marker */wa/* vs. object marker */ta/*. *Theory and Practice in Language Studies*, 3(7):1081, 2013.

[89] Radoslaw Kowalski, Marc Esteve, and Slava J. Mikhaylov. Application of natural language processing to determine user satisfaction in public services, 2017.

[90] Madura Kulatunga. Madura english-sinhala dictionary - online language translator. URL https://maduraonline.com/.

[91] M Anand Kumar, V Dhanalakshmi, KP Soman, and S Rajendran. Factored statistical machine translation system for english to tamil language. *Pertanika Journal of Social Sciences & Humanities*, 22(4), 2014.

[92] Peter M Landwehr and Kathleen M Carley. Social media in disaster relief. In *Data mining and knowledge discovery for big data*, pages 225–257. Springer, 2014. URL https://pdfs.semanticscholar.org/c666/41c6c15459186f492491df536fb9102db9d5.pdf.

[93] Jey Han Lau, David Newman, and Timothy Baldwin. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, 2014.

[94] William Li, David Larochelle, and Andrew Lo. Estimating policy trajectories during the financial crisis. 2014.

[95] William Li, Pablo Azar, David Larochelle, Phil Hill, and Andrew W Lo. Law is code: a software engineering approach to analyzing the united states code. *J. Bus. & Tech. L.*, 10:297, 2015.

[96] X. Li. The design and implementation of internet public opinion monitoring and analyzing system. In *2010 2nd International Conference on E-business and Information System Security*, pages 1–5, May 2010. doi: 10.1109/EBISS.2010.5473757.

[97] Elizabeth D Liddy. Natural language processing. 2001.

[98] Chamila Liyanage, Randil Pushpananda, Dulip Lakmal Herath, and Ruvan Weerasinghe. A computational grammar of sinhala. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 188–200. Springer, 2012.

[99] Jeevanthi Liyanapathirana and Ruvan Weerasinghe. English to sinhala machine translation: Towards better information access for sri lankans. In *Conference on Human Language Technology for Development*, pages 182–186, 2011.

[100] JU Liyanapathirana. A statistical approach to english and sinhala translation. 2013.

[101] R Loganathan. English-tamil machine translation system. *Master of Science by Research Thesis*, 2010.

[102] Sivaneasharajah Lushanthan, AR Weerasinghe, and DL Herath. Morphological analyzer and generator for tamil language. In *Advances in ICT for Emerging Regions (ICTer), 2014 International Conference on*, pages 190–196. IEEE, 2014.

[103] Prasenjit Majumder, Mandar Mitra, Dipasree Pal, Ayan Bandyopadhyay, Samaresh Maiti, Sukanya Mitra, Aparajita Sen, and Sukomal Pal. Text collections for fire. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 699–700. ACM, 2008.

[104] George Peiris Malalasekera. English-sinhalese dictionary. 1967.

[105] CS Malarkodi, RK Pattabhi, and Lalitha Devi Sobha. Tamil ner–coping with real time challenges. In *24th International Conference on Computational Linguistics*, page 23, 2012.

[106] SAPM Manamini, AF Ahamed, RAEC Rajapakshe, GHA Reemal, S Jayasena, GV Dias, and S Ranathunga. Ananya-a named-entity-recognition (ner) system for sinhala language. In *Moratuwa Engineering Research Conference (MERCon), 2016*, pages 30–35. IEEE, 2016.

[107] Inderjeet Mani and Mark T Maybury. Automatic summarization. 2001.

[108] William C Mann and Sandra A Thompson. *Rhetorical structure theory: A theory of text organization.* University of Southern California, Information Sciences Institute, 1987.

[109] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL http://www.aclweb.org/anthology/P/P14/P14-5010.

[110] Terrence Lance Martin. *Towards improved speech recognition for resource poor languages.* PhD thesis, Queensland University of Technology, 2006.

[111] Dr AG Menon, S Saravanan, R Loganathan, and Dr K Soman. Amrita morph analyzer and generator for tamil: a rule based approach. In *Proceedings of Tamil Internet Conference*, pages 239–243, 2009.

[112] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[113] David Mimno, Hanna M Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pages 880–889. Association for Computational Linguistics, 2009.

[114] T Mokanarangan, T Pranavan, U Megala, N Nilusija, Gihan Dias, Sanath Jayasena, and Surangika Ranathunga. Tamil morphological analyzer using support vector machines. In *International Conference on Applications of Natural Language to Information Systems*, pages 15–23. Springer, 2016.

[115] Prakash M Nadkarni, Lucila Ohno-Machado, and Wendy W Chapman. Natural language processing: an introduction. *Journal of the American Medical Informatics Association*, 18(5):544–551, 2011.

[116] Thilini Nadungodage and Ruvan Weerasinghe. Continuous sinhala speech recognizer. In *Conference on Human Language Technology for Development, Alexandria, Egypt*, pages 2–5, 2011.

[117] Thilini Nadungodage, Ruwan Weerasinghe, and Mahesan Niranjan. Speech recognition for low resourced languages: Efficient use of training data for sinhala speech recognition by active learning.

[118] Thilini Nadungodage, Ruvan Weerasinghe, and Mahesan Niranjan. Efficient use of training data for sinhala speech recognition using active learning. In *Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on*, pages 149–153. IEEE, 2013.

[119] Thilini Nadungodage, Ruvan Weerasinghe, and Mahesan Niranjan. Speaker adaptation applied to sinhala speech recognition. *Int. J. Comput. Linguistics Appl.*, 6(1):117–129, 2015.

[120] Lakshika Nanayakkara, Chamila Liyanage, Pubudu-Tharaka Viswakula, Thilini Nagungodage, Randil Pushpananda, and Ruvan Weerasinghe. A human quality text to speech system for sinhala. In *Proc. The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 157–161.

[121] Purnima Nanayakkara and Surangika Ranathunga. Clustering sinhala news articles using corpus-based similarity measures. In *2018 Moratuwa Engineering Research Conference (MERCon)*, pages 437–442. IEEE, 2018.

[122] Shrikanth Narayanan, Panayiotis G Georgiou, Abhinav Sethy, Dagen Wang, Murtaza Bulut, Shiva Sundaram, Emil Ettelaie, Sankaranarayanan Ananthakrishnan, Horacio Franco, Kristin Precoda, et al. Speech recognition engineering issues in speech to speech translation system design for low resource languages and domains. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 5. IEEE, 2006.

[123] John J Nay. Gov2vec: Learning distributed representations of institutions and their legal text. *arXiv preprint arXiv:1609.06616*, 2016.

[124] Ani Nenkova, Kathleen McKeown, et al. Automatic summarization. *Foundations and Trends® in Information Retrieval*, 5(2–3):103–233, 2011.

[125] Joel Larocca Neto, Alex A Freitas, and Celso AA Kaestner. Automatic text summarization using a machine learning approach. In *Brazilian Symposium on Artificial Intelligence*, pages 205–215. Springer, 2002.

[126] S. O'Halloran, S. Maskey, G. McAllister, D. K. Park, and K. Chen. Big data and the regulation of financial markets. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 1118–1124, Aug 2015. doi: 10.1145/2808797.2808841.

[127] Online Etymology Dictionary. Origin and meaning of human, 2018. URL https://www.etymonline.com/word/human.

[128] S Pandian, Krishnan Aravind Pavithra, and T Geetha. Hybrid three-stage named entity recognizer for tamil. *INFOS2008, March Cairo-Egypt. Available at: http://infos2008. fci. cu. edu. eg/infos/NLP_08_P045-052. pdf*, 2008.

[129] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135, 2008.

[130] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[131] K Parameshwari. An implementation of apertium morphological analyzer and generator for tamil. *Parsing in Indian Languages*, page 41, 2011.

[132] Braja Gopal Patra, Dipankar Das, Amitava Das, and Rajendra Prasath. Shared task on sentiment analysis in indian languages (sail) tweets-an overview. In *International Conference on Mining Intelligence and Knowledge Exploration*, pages 650–655. Springer, 2015.

[133] Randil Pushpananda, Ruvan Weerasinghe, and Mahesan Niranjan. Towards sinhala tamil machine translation. In *Advances in ICT for Emerging Regions (ICTer), 2013 International Conference on*, pages 288–288. IEEE, 2013.

[134] Randil Pushpananda, Ruvan Weerasinghe, and Mahesan Niranjan. Sinhala-tamil machine translation: Towards better translation quality. In *Proceedings of the Australasian Language Technology Association Workshop 2014*, pages 129–133, 2014.

[135] Dragomir R Radev. A common theory of information fusion from multiple text sources step one: cross-document structure. In *Proceedings of the 1st SIGdial workshop on Discourse and dialogue-Volume 10*, pages 74–83. Association for Computational Linguistics, 2000.

[136] K Rajan, Vennila Ramalingam, M Ganesan, S Palanivel, and B Palaniappan. Automatic classification of tamil documents using vector space model and artificial neural network. *Expert Systems with Applications*, 36(8):10914–10918, 2009.

[137] S Rajendran. Preliminaries to the preparation of a word net for tamil. *Language in India*, 2(1), 2002.

[138] S Rajendran. Parsing in tamil: Present state of art. *Language in India*, 6:8, 2006.

[139] S Rajendran, S Arulmozi, B Kumara Shanmugam, S Baskaran, and S Thiagarajan. Tamil wordnet. In *Proceedings of the First International Global WordNet Conference. Mysore*, volume 152, pages 271–274, 2002.

[140] Loganathan Ramasamy and Zdeněk Žabokrtskỳ. Tamil dependency parsing: results using rule based and corpus based approaches. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 82–95. Springer, 2011.

[141] Loganathan Ramasamy, OndÅ ej Bojar, and ZdenÄ k Å1/2abokrtskÃ1/2. Morphological processing for english-tamil statistical machine translation. In *Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages*, pages 113–122, 2012.

[142] Surangika Ranathunga, Fathima Farhath, Uthayasanker Thayasivam, Sanath Jayasena, and Gihan Dias. Si-ta: Machine translation of sinhala and tamil official documents. In *2018 National Information Technology Conference (NITC)*, pages 1–6. IEEE, 2018.

[143] Uma Maheshwar Rao and K Prameshwari. On the description of morphological data for morphological analyzers and generators: A case study of telgu tamil & kannada. *Knowledge Sharing Events. LDC-IL, CIIL, Mysore*, 2010.

[144] Prasad Rashmi, Dinesh Nihkil, Lee Alan, Miltsakaki Eleni, Robaldo Livio, Joshi Aravind, Webber Bonnie, et al. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco, may. European Language Resources Association (ELRA).*, 2008.

[145] Gathika Ratnayaka, Thejan Rupasinghe, Nisansa de Silva, Menuka Warushavithana, Viraj Gamage, and Amal Shehan Perera. Identifying relationships among sentences in court case transcripts using discourse relations. In *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 13–20. IEEE, 2018.

[146] Raghunathan S. Ravishankar, N. Corpus based sentiment classification of tamil movie tweets using syntactic patterns. pages 172–178, 2017.

[147] Vasu Renganathan. Development of part-of-speech tagger for tamil. In *Tamil Internet 2001 conference*, 2001.

[148] Vasu Renganathan. An interactive approach to development of english-tamil machine translation system on the web. In *The international Tamil Internet 2002 Conference and Exhibition (TI2002)*, 2002.

[149] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Pearson Education Limited,, 2016.

[150] Navanath Saharia, Utpal Sharma, and Jugal Kalita. Stemming resource-poor indian languages. 13(3): 14:1–14:26, October 2014. ISSN 1530-0226. doi: 10.1145/2629670. URL http://doi.acm.org/10.1145/2629670.

[151] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pages 851–860. ACM, 2010.

[152] Richard Salomon. *Indian epigraphy: a guide to the study of inscriptions in Sanskrit, Prakrit, and the other Indo-Aryan languages*. Oxford University Press, 1998.

[153] Karen Sanders and María José Canel. Mind the gap: Local government communication strategies and spanish citizens' perceptions of their cities. *Public Relations Review*, 41(5):777–784, 2015.

[154] Edward Sapir. Chapter 7. language as a historical product: Drift. edward sapir. 1921. language: An introduction to the study of speech, 1921. URL https://www.bartleby.com/186/7.html.

[155] Sunita Sarawagi et al. Information extraction. *Foundations and Trends® in Databases*, 1(3):261–377, 2008.

[156] Michael Schmitz, Robert Bart, Stephen Soderland, Oren Etzioni, et al. Open language learning for information extraction. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 523–534. Association for Computational Linguistics, 2012.

[157] Richard Schwartz, Toru Imai, Francis Kubala, Long Nguyen, and John Makhoul. A maximum likelihood model for topic classification of broadcast news. In *Fifth European Conference on Speech Communication and Technology*, 1997.

[158] Shriya Se, R Vinayakumar, M Anand Kumar, and KP Soman. Predicting the sentimental reviews in tamil movie using machine learning algorithms. *Indian Journal of Science and Technology*, 9(45), 2016.

[159] M Selvam and AM Natarajan. Improvement of rule based morphological analysis and pos tagging in tamil language via projection and induction techniques. *International journal of computers*, 3(4):357–367, 2009.

[160] Anutharsha Selvarasa, Nilasini Thirunavukkarasu, Niveathika Rajendran, Chinthoorie Yogalingam, Surangika Ranathunga, and Gihan Dias. Short tamil sentence similarity calculation using knowledge-based and corpus-based similarity measures. In *Engineering Research Conference (MERCon), 2017 Moratuwa*, pages 443–448. IEEE, 2017.

[161] A Sharmista and M Ramaswami. Tree based opinion mining in tamil for product recommendations using r,“, 2016.

[162] Buddhika Siddhisena. Firefox තුලට ඉංග්‍රීසි-සිංහල ශබ්ද කෝෂයක්. URL http://www.sinhalenfoss.org/2008/02/ensitip/.

[163] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[164] KG Srinivasagan, S Suganthi, and N Jeyashenbagavalli. An automated system for tamil named entity recognition using hybrid approach. In *2014 International Conference on Intelligent Computing Applications (ICICA)*, pages 435–439. IEEE, 2014.

[165] Sakthithasan Sripirakas, AR Weerasinghe, and Dulip L Herath. Statistical machine translation of systems for sinhala-tamil. In *Advances in ICT for Emerging Regions (ICTer), 2010 International Conference on*, pages 62–68. IEEE, 2010.

[166] Keet Sugathadasa, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, and Madhavi Perera. Synergistic union of word2vec and lexicon for domain specific semantic similarity. *IEEE International Conference on Industrial and Information Systems (ICIIS)*, 2017.

[167] Keet Sugathadasa, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, and Madhavi Perera. Legal document retrieval using document vector embeddings and deep learning. *arXiv preprint arXiv:1805.10685*, 2018.

[168] Oscar Tay. What is the etymology of the word "human"?, 2018. URL https://goo.gl/sAuoUp.

[169] Oscar Tay. Why does 'woman' contain 'man' and 'female' contain 'male'?, 2018. URL https://goo.gl/dkp4Cs.

[170] Pasindu Tennage, Prabath Sandaruwan, Malith Thilakarathne, Achini Herath, Surangika Ranathunga, Sanath Jayasena, and Gihan Dias. Neural machine translation for sinhala and tamil languages. In *Asian Language Processing (IALP), 2017 International Conference on*, pages 189–192. IEEE, 2017.

[171] Pasindu Tennage, Achini Herath, Malith Thilakarathne, Prabath Sandaruwan, and Surangika Ranathunga. Transliteration and byte pair encoding to improve tamil to sinhala neural machine translation. In *2018 Moratuwa Engineering Research Conference (MERCon)*, pages 390–395. IEEE, 2018.

[172] Pasindu Tennage, Prabath Sandaruwan, Malith Thilakarathne, Achini Herath, and Surangika Ranathunga. Handling rare word problem using synthetic training data for sinhala and tamil neural machine translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.

[173] Horia-Nicolai Teodorescu. Using analytics and social media for monitoring and mitigation of social disasters. *Procedia Engineering*, 107:325 – 334, 2015. ISSN 1877-7058. doi: https://doi.org/10.1016/j.proeng.2015.06.088. URL http://www.sciencedirect.com/science/article/pii/S1877705815010413. Humanitarian Technology: Science, Systems and Global Impact 2015, HumTech2015.

[174] Mokanarangan Thayaparan, Surangika Ranathunga, and Uthayasanker Thayasivam. Graph based semi-supervised learning approach for tamil pos tagging. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.

[175] Pranavan Theivendiram, Megala Uthayakumar, Nilusija Nadarasamoorthy, Mokanarangan Thayaparan, Sanath Jayasena, Gihan Dias, and Surangika Ranathunga. Named-entity-recognition (ner) for tamil language using margin-infused relaxed algorithm (mira). In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 465–476. Springer, 2016.

[176] Suhatati Tjandra, Amelia Alexandra Putri Warsito, and Judi Prajetno Sugiono. Determining citizen complaints to the appropriate government departments using knn algorithm. In *ICT and Knowledge Engineering (ICT & Knowledge Engineering 2015), 2015 13th International Conference on*, pages 1–4. IEEE, 2015.

[177] Andrew T Tsubaki. Zeami and the transition of the concept of yūgen: A note on japanese aesthetics. *Journal of Aesthetics and Art Criticism*, pages 55–67, 1971.

[178] Alan M Turing. Computing machinery and intelligence. In *MIND: A Quarterly Review of Psychology and Philosophy*, volume 59, pages 433–460. 1950.

[179] University of Pennsylvania. Iiit tagset guidelines. URL http://shiva.iiit.ac.in/SPSAL2007/iiit_tagset_guidelines.

[180] Dimuthu Upeksha, Chamila Wijayarathna, Maduranga Siriwardena, Lahiru Lasandun, Chinthana Wimalasuriya, N. H. N. D. De Silva, and Gihan Dias. Implementing a Corpus for Sinhala Language. In *Symposium on Language Technology for South Asia 2015*, 2015.

[181] R Vijayakrishna and L Sobha. Domain focused named entity recognizer for tamil using conditional random fields. In *Proceedings of the IJCNLP-08 Workshop on Named Entity Recognition for South and South East Asian Languages*, 2008.

[182] Sonny Vu, Christopher Bader, and David Purdy. Automatic summarization of a document, June 20 2002. US Patent App. 09/908,443.

[183] Hanna M Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM, 2006.

[184] Chong Wang and David M Blei. Collaborative topic modeling for recommending scientific articles. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 448–456. ACM, 2011.

[185] Asanka Wasala and Ruvan Weerasinghe. Ensitip: a tool to unlock the english web. In *11th international conference on humans and computers, Nagaoka University of Technology, Japan*, pages 20–23, 2008.

[186] Amali Weerasinghe and Gihan Dias. Construction of a multilingual place name database for sri lanka. 2013.

[187] Ruvan Weerasinghe. A statistical machine translation approach to sinhala-tamil language translation. *Towards an ICT enabled Society*, page 136, 2003.

[188] Ruvan Weerasinghe, Asanka Wasala, Viraj Welgama, and Kumudu Gamage. Festival-si: A sinhala text-to-speech system. In *International Conference on Text, Speech and Dialogue*, pages 472–479. Springer, 2007.

[189] Viraj Welgama, Dulip Lakmal Herath, Chamila Liyanage, Namal Udalamatta, Ruvan Weerasinghe, and Tissa Jayawardana. Towards a sinhala wordnet. In *Proceedings of the Conference on Human Language Technology for Development*, 2011.

[190] Viraj Welgama, Ruvan Weerasinghe, and Mahesan Niranjan. Evaluating a machine learning approach to sinhala morphological analysis. In *Proceedings of the 10th International Conference on Natural Language Processing, Noida, India*, 2013.

[191] Yudhanjaya Wijeratne. The control of hate speech on social media: Lessons from sri lanka. *CPR South*, 2018.

[192] Indeewari Wijesiri, Malaka Gallage, Buddhika Gunathilaka, Madhuranga Lakjeewa, Daya Wimalasuriya, Gihan Dias, Rohini Paranavithana, and Nisansa De Silva. Building a wordnet for sinhala. In *Proceedings of the Seventh Global WordNet Conference*, pages 100–108, 2014.

[193] Daya C Wimalasuriya and Dejing Dou. Ontology-based information extraction: An introduction and a survey of current approaches, 2010.

[194] Miron Winslow and Joseph Knight. *A comprehensive Tamil and English dictionary of high and low Tamil*. PR Hunt, 1862.

[195] Florian Wolf, Edward Gibson, Amy Fisher, and Meredith Knight. Discourse graphbank. *Linguistic Data Consortium, Philadelphia*, 2004.

[196] Chang Wu and Yidong Chen. A survey of researches on the application of natural language processing in internet public opinion monitor. In *2011 International Conference on Computer Science and Service System (CSSS)*, pages 1035–1038, June 2011. doi: 10.1109/CSSS.2011.5972059.

[197] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics, 1994.

[198] Holly Young. A language family tree - in pictures | education | the guardian. URL https://www.theguardian.com/education/gallery/2015/jan/23/a-language-family-tree-in-pictures.