Linked and Structured Tamil Data for Machine Learning

Dr. Saatviga Sudhahar
Machine Learning Scientist

Linked Data for Machine Learning

What is Linked Data?

- Linked data is structured data which is interlinked with other data so it becomes more useful through semantic queries.
- It builds upon standard Web technologies such as HTTP, RDF and URIs, but rather than using them to serve web pages only for human readers, it extends them to share information in a way that can be read automatically by computers.
- Linked data may also be open data, in which case it is usually described as linked open data (LOD).

Linked data standards

- Use URIs to name (identify) things.
- Use HTTP URIs so that these things can be looked up (interpreted, "dereferenced").
- Provide useful information about what a name identifies when it's looked up, using open standards such as RDF, SPARQL, etc.

Example Linked databases

- DBpedia – a dataset containing extracted data from Wikipedia; it contains about 3.4 million concepts described by 1 billion triples, including abstracts in 11 different languages
- FOAF – a dataset describing persons, their properties and relationships
- GeoNames – provides RDF descriptions of more than 7,500,000 geographical features worldwide.
- UMBEL – a lightweight reference structure of 20,000 subject concept classes and their relationships derived from OpenCyc, which can act as binding classes to external data; also has links to 1.5 million named entities from DBpedia and YAGO
- Wikidata – a collaboratively-created linked dataset that acts as central storage for the structured data of its Wikimedia Foundation sister projects

Data analysis

- Linked data allows users to explore data in more novel ways.
    - Answering complex questions
        - Eg: Give me the names of all pianists taught by x, where x was taught the piano by Liszt
        - Requires building ontologies
    - Linked data is used to build knowledge graphs that connect information about entities of various types. Eg: Google’s knowledge graph.
    - Allowing to answer questions over a knowledge graph.
    - Infer links to use for recommendation over knowledge graphs
    - Knowledge graph completion techniques with machine learning help in adding more links to the data by enriching it

Linked data for Machine Learning

- Machine learning requires training models with large amounts data.
- The availability of more features increases the learning capacity of the model.
- With linked data it is possible to,
    - discover automatically datasets where the entities of interest occur
    - show to the user a big number of useful features for these entities and
    - create automatically the selected features by sending SPARQL queries.

Structured data for Machine Learning

Useful applications

- Spell checker
- Tamil paraphrase detection
- Tamil speech to text and text to speech systems
- Machine translation systems - English to Tamil, Tamil to English, Tamil to Dravidian languages
- Information extraction insights
    - Detecting historical trends for topics
    - Sentiment analysis of overall text, specific topics in text, sentiment timelines
- Grammar checker
- Question answering systems

Natural language processing for Tamil

- Processing of Tamil language for machine learning involves contribution in the field of Tamil natural language processing which is to process and analyze large amounts of natural language data.
- This involves work in several sub areas including,
  a. Building corpus text to train models (significant for machine learning with Tamil text)
     i. Monolingual corpus
     ii. Parallel corpus - used in machine translation
     iii. Annotated corpus(POS tagged)
     iv. Speech Corpora

a. Building Language models

  i. important resource for various NLP applications that require generation of text including Machine Translation, Speech Recognition etc.

b. Lexical dictionaries

a. Tools for language parsing and resolution

  i. Tokeniser - Tokenization is the process of breaking a stream of text up into words, phrases, symbols, or other meaningful elements called tokens.

  ii. Chunker - Chunking is the task of identifying and segmenting the text into syntactically correlated word groups. It could divide a sentence into its major non-overlapping phrases and attach a label to each chunk.

  iii. sentence splitting - Sentence boundary disambiguation (SBD), also known as sentence breaking, is the problem in natural language processing of deciding where sentences begin and end.

  iv. part-of-speech tagging - Part of speech (POS) tagging is the process of labeling a part of speech or other lexical class marker to each and every word in a sentence.

v. dependency parsing - Dependency parsing uses linguistic information to give relationship between words.

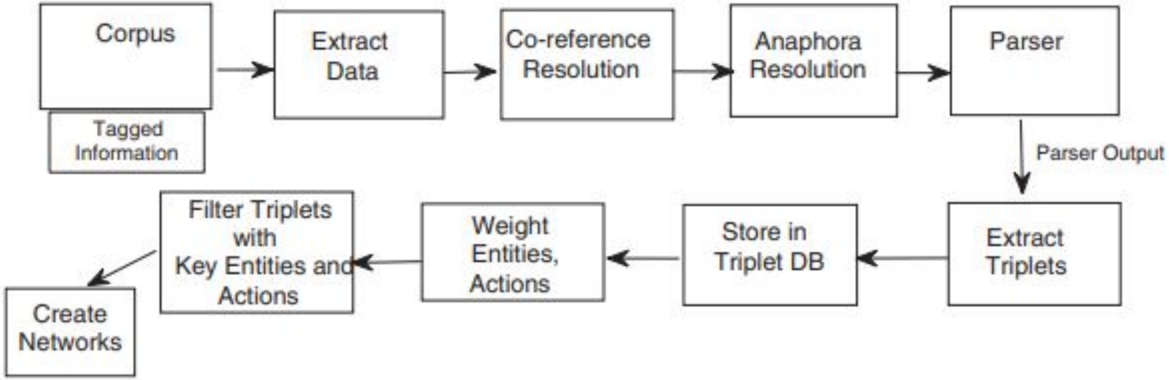
vi. coreference/anaphora resolution - This analysis finds the nouns and phrases that refer to the same entity enabling the extraction of relations among entities as well as more complex propositions.

vii. named entity recognition - Named Entity Recognition is the process of identifying and recognizing named entities such as person, organization, location, date, time and money in the text documents.

viii. Word sense disambiguation - Word sense disambiguation is the process of identifying the right sense for a word when a word might have two or more meanings. This is vital for problems such as machine translation, question and answering or information retrieval.

Large scale textual data analysis

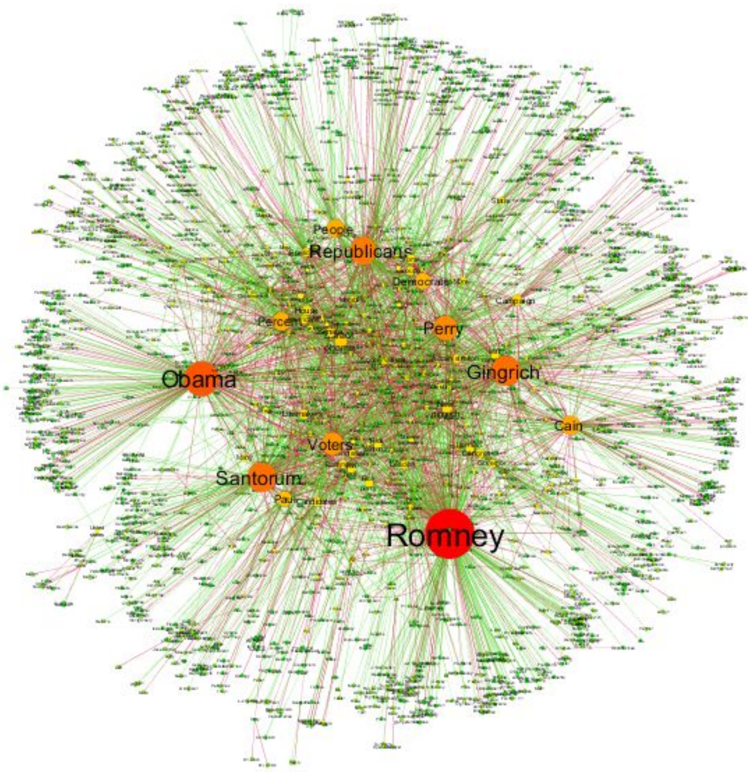
- Turn a corpus into a network of actors, objects and actions.
- The software pipelines used to extract the information out of a large text corpus of news data
- News data was collected on a daily basis by crawling news websites and using topic classifiers to classify and tag them in different topics. Eg: education, entertainment, sports, elections etc.



Large scale textual data analysis

- By weighting the actors, we can identify the players most identified with a given domain (eg: crime);
- by analysing the centrality of the actors, we can identify the most influential characters in the news narrative;
- by classifying the types of actions (eg crimes against person) we can further analyse the roles different actors play in crime (eg: perpetrator vs. victim);
- by analysing the time series of the actors centrality, we can identify important changes in its narrative role (as done by hand in (Franzosi (2010)) where the emergence of Italian fascism was investigated in the same way).

Narrative networks showing patterns/communities



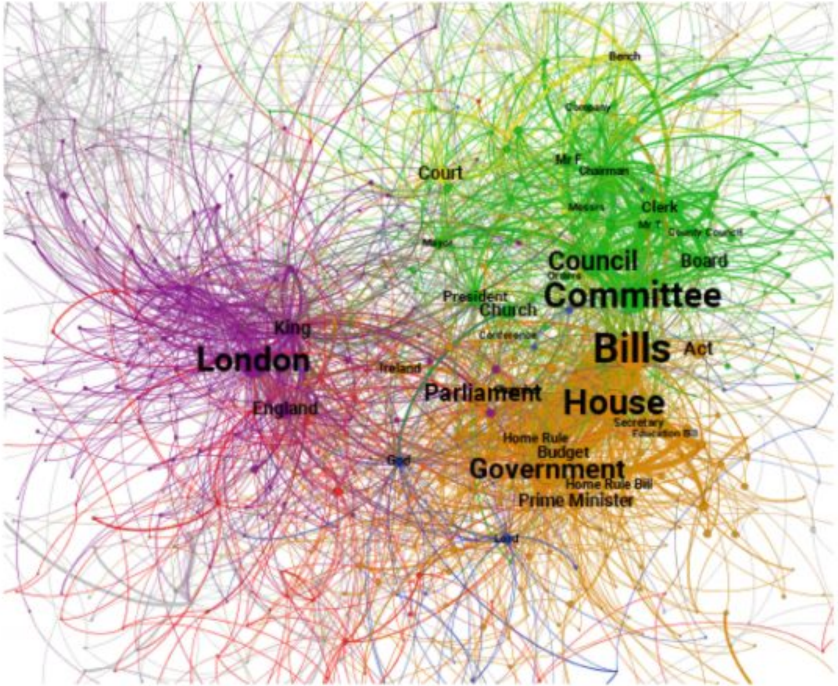


Fig. 1. Narrative network of the actors (nodes) and actions (edges) performed by them between 1905 and 1915 in the British newspaper corpus. Nodes are coloured based upon the community to which they belong.

Identified gaps for Tamil language for text analysis/ML models

- Non-availability of large and structured/annotated text corpus
  - With CTNLPR we aim to collect data, structure and curate them
- Several tools built for Tamil NLP, but it’s difficult to evaluate their usefulness
- There is room to improve the accuracy of these systems but the bottleneck lies in the availability of data to train such accurate systems.
- Scalability issues in running Tamil NLP tools
- Interpreting, tuning ML models requires more deeper analysis of features that need to be extracted by the tools

Thank You