TEQIP - II / World Bank Sponsored Centre of Excellence

KAnOE - Centre for Knowledge Analytics and Ontological Engineering

PES Institute of Technology / PES University

"PES Ullekha Maalaa"

One Million Kannada Wiki Pages

http://kanoe.org/kannada.html

Dr. Kavi Mahesh

Developed by Mr. Anantha Krishna Thantri, MTech

with contributions from Mr. Udaya Kiran P., MTech

November 2014

The Internet is a major source of information for many of us. However, much of the information on the Internet is available only in English and a few other Western European languages. Speakers of other languages, especially Bharatheeya languages such as Kannada, are unable to benefit from the vast resources on the net. For example, while English Wikipedia (http://en.wikipedia.org) has about 46,43,000 pages, Kannada Wikipedia (http://kn.wikipedia.org) has only 17,000 pages, or a mere 0.37% of English, after being in existence for a dozen years. Other local sources such as the KaNaja (http://kanaja.in) are also rather small; KaNaja has about 13,000 pages.

Although decades of research has gone into developing Machine Translation, it is yet to deliver usable technology, especially when it comes to translating from English to a rather dissimilar language such as Kannada. Google Translator, for example, is not accurate enough to be useful at this time, to automatically convert English articles in Wikipedia to Kannada.

Ullekha Maalaa was created to address this problem by developing a radically new approach to generating information pages in Kannada [Mahesh and Thantri, in preparation]. The idea was to start with information, not with English pages at all. Highly accurate datasets such as *YAGO* and *Dbpedia* are freely available. *YAGO - Yet Another Great Ontology* - is a dataset extracted from Wikipedia by *Max Plank Institute for Informatics* in Germany. It has about 19,00,00,000 (19 crore) facts about people, places, countries, songs, movies, etc. Much of the information in *Ullekha Maalaa* was taken from *YAGO*.

Facts in *YAGO* were converted to Kannada to create the world's first *Linked Open Dataset* in Kannada - also the first in any Bharatheeya language, to the best of our knowledge. This was done by transliterating all the names and values in the data (about 2.5 crore of them) while manually translating a small number (less than a hundred) of meaningful words. For example, *wasBornOnDate* was translated as "... ru ... andu janisidaru." (“ .. ರು .. ಅಂದು ಜನಿಸಿದರು.”)

The Kannada dataset has 2,50,00,000 (2.5 crore) pieces of information about 1,00,00,000 (one crore) different subjects. From these facts, an equal number of Kannada sentences (i.e., 2.5 crore) were generated. Natural Language Generation techniques were then applied to the set of sentences to generate documents. Related sentences were aggregated to generate longer sentences. The sentences were grouped into paragraphs and arranged in proper order to generate the flow of the document by applying an ontology of predicates present in the dataset. This idea of using an ontology to order predicates is also an original contribution of the research work carried out at KAnOE [Mahesh et al., 2012, 2013, Chari et. al., 2014].

After filtering out very short documents, we were left with one million (10 lakh) Kannada documents in *Ullekha Maalaa*. Almost every stage in this process required very heavy-duty computing which took several days or weeks to execute even on the state-of-the-art computing servers available at KAnOE (servers with 2 CPU, 12 core, 96 GB RAM and 8x500 GB RAID disks). For example, replacing and correcting about 800 misspelled Kannada words in the one million documents took about 100 Billion string matching operations (i.e., 100,000,000,000 or 10,000 crore string comparisons).

The Kannada pages in *Ullekha Maalaa* are by no means complete or correct. However, they have been made freely available on our web site (hosted on an Amazon cloud server) and can be edited by anyone, directly in the browser, *without using any special software* (i.e., no need to install *nudi, lipi,* or *baraha*). *Ullekha Maalaa* is available at http://kanoe.org/kannada.html

Here is hoping that a large number of Kannadigas will enthusiastically participate in editing and improving the quality of language and the quantity of information in *Ullekha Maalaa* to make it a very useful free resource of information for Kannadigas everywhere.

About KAnOE: KAnOE - the Centre for Knowledge Analytics and Ontological Engineering (http://kanoe.org) - is a research centre at PES University funded primarily by the World Bank through the Government of India - Ministry of Human Resources Development's TEQIP program, along with the Government of Karnataka. With a focus on the unique area of combining analytics with ontological engineering, KAnOE aims to make a defining contribution to the field of knowledge management through knowledge analytics and ontological engineering. KAnOE has been funded also by PES University, AICTE through its MODROBS program, Department of Electronics and Information Technology (Ministry of Communication and Information Technology) through its TDIL program, Visvesvaraya Technological University and Microsoft Research-India. KAnOE has Billion Data Capability, i.e., to process and index a billion (100 crore) pieces of information in just about two hours using state-of-the-art computer hardware and graph database software.

About PESIT / PES University: PES University (http://www.pes.edu) is a leading State Private University in Bangalore. Formerly known as PES Institute of Technology, it has been a top-ranked institution in the country in engineering and management studies for the last 25 years. With a strong focus on research, PES University aims to become one of the best research universities in India. It has several Centres of Excellence and identified domains of research apart from the interdisciplinary Crucible of Research and Innovation.

References:

• [Mahesh et al., 2012] Kavi Mahesh, Pallavi Karanth (2012) Smart-Aleck: An Interestingness Algorithm for Large Semantic Datasets In: 11th International Semantic Web Conference - Semantic Web Challenge, 11-15 Nov, 2012, Boston, MA, USA.

• [Mahesh et al., 2013] Kavi Mahesh, Shruthi Chari, Shrinidhi Ramakrishnan (2013) LODScape: Ontology-Based Multiple-LOD Object Browser In: Proc. 12th International Semantic Web Conference, ISWC-2013, Semantic Web Challenge, 21-25 October 2013, Sydney, Australia.

• [Chari et al., 2014] Shruthi Chari, Shrinidhi Ramakrishnan, Kavi Mahesh (2014) LODMedics: Bringing Semantic Data to the Common Man In: Proc. 13th International Semantic Web Conference, ISWC-2014, Semantic Web Challenge, 21-23 October 2014, Riva del Garda, Trentino, Italy.

• [Mahesh and Thantri, in preparation] Kavi Mahesh, Anantha Krishna Thantri (in preparation) Machine Translation from Linked Open Datasets.