Open Domain Information Extraction with Harmonic Energy Functions

Chirag Nagpal\textsuperscript{1}, Kyle Miller and Artur Dubrawski\textsuperscript{2}

\textsuperscript{1}Language Technologies Institute, \textsuperscript{2}Auton Lab, Robotics Institute, Carnegie Mellon University

\{chiragn, mille856, awd\} @ andrew.cmu.edu

Objectives

We build upon the Harmonic Energy Functions used previously in an Active Learning Framework

\begin{itemize}
  \item Named Entities extracted with Distant Supervision from external Knowledge Bases are exploited to describe a Graph with weighted edges.
  \item Entities to provide context, and constrain search to utilise Active Search for Information Extraction from Open Domain data.
  \item Scalability issues arising due to large number of nodes are addressed with numerical techniques, that speed solution of Linear Systems.
\end{itemize}

Introduction

To extract entities, we exploit external KBs:

\begin{itemize}
  \item Wordnet: Named Entity Recogniser and Part of Speech Tagger trained on the Penn Treebank Corpus is employed along with Lesk, to perform Word Sense Disambiguation.
  \item DBPedia: 'Spotlight'\textsuperscript{[2]} an open-source tool to map text to Entities.
\end{itemize}

Architecture

To extract entities, we exploit external KBs:

\begin{itemize}
  \item Wordnet: Named Entity Recogniser and Part of Speech Tagger trained on the Penn Treebank Corpus is employed along with Lesk, to perform Word Sense Disambiguation.
  \item DBPedia: 'Spotlight'\textsuperscript{[2]} an open-source tool to map text to Entities.
\end{itemize}

Information Extraction

\begin{itemize}
  \item Energy Minimisation is extended towards a more Generic Open Domain Information Extraction System
  \item A structured dataset curated from Wikipedia articles of countries is utilised to answer queries such as ‘Arabic Speaking African Nations’, ‘Most Democratic Nations’.
\end{itemize}

Figure 1: In each iteration, the a label associated with the node is minimised the Energy over the Graph.

Figure 2: Data Preprocessing Pipeline

Active Search, is applied to 20 Newsgroups for a One-Vs-Rest Classification task.

AS on Bi-Partite graphs with Entities significantly outperforms AS on Graphs with Cosine Distance between nodes.

Figure 3: Comparative Entities Extracted with WordNet and DBPedia on a representative sample.

Future Work

\begin{itemize}
  \item Treat change in \( f(t+1) \) w.r.t. \( f(t) \) as Impact Factor (IF).
  \item Faster computation given IF is a function of \( f \) or \( f^T f \).
  \item Investigate use of a classifier, to learn IF.
\end{itemize}

Figure 4: Harmonic Energy Minimisation - Active Search

References