Chapter 18
Effective Entity Linking and Disambiguation Algorithms for User-Generated Content (UGC)

Senthil Kumar Narayanasamy
VIT University, India

Dinakaran Muruganantham
VIT University, India

ABSTRACT

The exponential growth of data emerging out of social media is causing challenges in decision-making systems and poses a critical hindrance in searching for the potential information. The major objective of this chapter is to convert the unstructured data in social media into the meaningful structure format, which in return brings the robustness to the information extraction process. Further, it has the inherent capability to prune for named entities from the unstructured data and store the entities into the knowledge base for important facts. In this chapter, the authors explain the methods to identify all the critical interpretations taken over to find the named entities from Twitter streams and the techniques to proportionally link it with appropriate knowledge sources such as DBpedia.

INTRODUCTION

The conventional methods followed for information extraction in text documents (as they are structured and well-formed) is totally different with information extraction in social media contents. The social media contents are mostly unstructured and especially ill-formed to extract the information from it. As stated by the authors Laere, Schockaert, Tanasescu, Dhoedt, & Jones (2014) and Giridhar, Abdelzaher, George, & Kaplan (2015, March), it was estimated that the accuracy rate of precision in structured documents is pointing to 89% whereas unstructured documents hold below 64%. To culminate this difference, several approaches have been discussed and techniques were proposed to boost the precision and recall rate of unstructured documents as given by Lee, Ganti, Srivatsa, & Li (2014, December) and Imran, Castillo, Diaz, & Vieweg (2015); but still, problems persist and pertaining in many situations. In

DOI: 10.4018/978-1-5225-5384-7.ch018

order to streamline the accuracy rate over precision and recall, we have here proposed some methods to augment the precision and use new strategies to overcome the impeding difficulties.

To start with the extraction process, the principal task is to find the potential named entities out from the unstructured text. In our case, we have taken Twitter social media content and identified the named entities from its streams. But the objectivity comes when we deal with real world entities which have been mapped with one-to-many cardinality over knowledge sources and pinches in for the major setbacks for further processes. Besides as the tweets are very short and most of the instances informal in nature, finding potential named entities out of tweet is a crucial task for any automated systems. This sort of ambiguity conundrum is very high in information retrieval context and yields huge difficulties to Named Entity Recognition (NER) systems. To conduct entity identification process, we have used the Markov Network (Lee et al., 2014, December), that was deployed for many conventional information extraction tasks and yielded high accuracy rate. In our cases as we have taken Twitter social media streams, the entities were represented with nodes and the edges will get connected between the conditional dependencies over selected named entities. If we dig deep closer to this whole network, it would almost resemble to Bayesian Network except the fact that edges were cyclic and undirected. For any document, the entity is appropriately mapped with its sheer interpretation of selected named entities suggested by the knowledge source. In some worst cases as we had witnessed in the empirical results, it has shown that few entities has no link to relate with the knowledge source and it has paved way for ambiguous connection and lead to bad search results. This was taken as one of the research gap identified in the extraction process and we had given the solution for the same in the following sections.

The Hidden Markov Model uses many language processing tasks such as POS tagging, Named Entity Detection, and Classification, etc. In this proposed approach, we have taken Twitter as a social media site and carry out the process of identifying the potential named entities from Twitter streams. As the tweets are very short and noisy, finding named entities is a challenging task and linking named entities to appropriate knowledge base mentions is yet another cumbersome process to deal with. Hence, in this proposed system, we have explained the mechanism to link entities to knowledge base, removing the ambiguity persisting over the extracted named entities and enhance the capabilities of searching much easier than before using semantic Web technologies like RDF/SPARQL.

Linking to Web Content

This part of the framework manages extraction of all the entities identified with the entered term from Web. This can comprehensively incorporate Wikipedia pages, news articles, blog entries, and so on. To catch a feeling of all the entities identified with the entered search term, the procedure is done in two levels. At the primary level there is an ordinary ‘Google Search’ through the programming interface. For example, R programming utilizes the ‘getGoogleURL’ and ‘getGoogleLinks’ calls to scan for the given subject. It impersonates an ordinary Web search however specifically posts the list items on the R interface. These outcomes frame the premise of the rundown. At the second level of handling a few connections are chosen from the Google list items and the entities on these site pages are scratched for synopsis. The RCurl bundle in R for writing computer programs is one of the effectively accessible bundles that permit content extraction upon providing it with a URL. Consequently, at this level of handling, we get all the Web content accessible on the World Wide Web relating to the point.

The essential partner of engineering is the potential users of the system. The user enters the query for acquiring search results and consequently is the essential hotspot for the information. The user is