Analysing Narratives: Automatic Descriptive Feature Extraction Through Latent Entity modelling

Adam Slack

Contents

Chapter 1

Introduction

1.1 Introduction

The Web; A seemingly ever-expanding resource, with data being generated and information published at an accelerating rate [1]. As more gain access to the internet, the rate at which new information is made available will only increase. Whether the origins of data be, social media, news outlets, or e-commerce reviews, much of the resources on the web exist in a natural language format. From this data there exists underlying information that can be extracted and utilised in decision making process. Given the amount of processing required to consume data on this scale, it is necessary to ensure that computational methods exist that can accommodate for individual or business needs to understand data. Many methods for Methods that allow the processing of data to extract surface level information exist, however there is room for further understanding. By relating distinct pieces of information or subsets of data, it is possible to frame or contextualise a solution to a problem within a wider setting. This paper aims to provide a novel method capable of providing succinct summaries of documents in terms of entity topics.

Information Retrieval (IR), Machine Learning (ML), and Natural Language Processing (NLP) when considered in conjunction with each other, concern them-

selves with the extracting of information from data in a natural language texts. Relating textual data yields information valuable to many different entities, including businesses when marketing products, individuals when choosing what to read, and even researchers considering the current state of a research area. For a business, relating entities extracted from texts, can help identify target groups to aim products at. For an individual, relating entities can assist in decisions on what to read, buy, and watch based on similarities between things that they do and don't like. For research, the identification of common themes, or prominent authors can be achieved through the relating of entities.

A Latent Entity in this investigation is defined as an abstract representation of some entity that can be used to describe one or more concrete entities. Thus the term Latent Entity Modelling is defined as a transformation from a corpus of text to a collection of Latent Entities describing a corpus. Latent Entities represent a layer of abstraction from a corpus, wherein it is possible to describe the corpus as a whole in terms of the Latent Entities. Entity models build upon the concept of topic modelling, particularly the kinds of topics derived through Latent Dirichlet Allocation. A Topic in this sense is a probability distribution of words, such that the degree of membership for each word in a topic indicates the probability of that word being an indicator of that topic.

1.2 Aims and Objectives

The aim of this project the value of information that can be extracted from text through the use of Latent Entity Modelling

1.3 Motivation

Motivations for this project include the need for more effective...

Chapter 2

Literature Review

2.1 Introduction

Natural Language Processing (NLP) is a vast field of study within Computer Science and Computational Intelligence, the focus of which is the development of intelligent systems capable of handling data in the form of natural language. The task of extracting Latent Entity Models is one that will touch upon many areas of NLP, including Part-of-Speech (POS) tagging, Named Entity Recognition (NER), and Latent Topic Modelling. Given that the focus of this paper is the extraction of descriptive information of narratives, it is necessary to consider the field of Information Retrieval and existing methods of descriptive summarisation as well.

It is possible to divide NLP into two large schools of thought; One involves the processing of natural language such that machine usable resources are available, the other is concerned with the application of information resulting from processed natural language. Whilst many NLP tasks require methods for parsing, understanding, or synthesising spoken words, this paper is only considering written texts. As such, the spoken language aspects of NLP will be overlooked.

Parsing written texts begins with tokenization. This is the division of a single string of characters in to strings representing sentences or words. Often text is tokenized into sentences, and each sentence is then tokenized into words. Once

a text is expressed with basic sentence and word structure it is possible to apply additional processing steps. POS tagging is the application of tags to sequences of words, each tag and resulting sequence represents the grammatical structure of a sequence of words. NER is the extraction of entities from natural language, entities are tagged depending on their type. NER is performed after tokenization, however it doesn't necessarily rely on text being POS tagged first.

There are a range of NLP tasks involving the applications of natural language parsing. From Customer relation Chatbots, to Document Summarisation and Relation. Whilst many applications involve the use of parsed text features, many also orient themselves around additional features. For example, tools for relating and summarising distinct documents may utilise topic models and topic modelling techniques.

2.2 Part-Of-Speech Tagging

2.2 Part-Of-Speech Tagging POS Tagging, is a useful task for many NLP problems. It forms a solid platform from which many investigations can be launched. The quality of a POS tagger can make or break a study. There is a range of POS tagging methods to choose from, many the highest performing taggers employ Maximum Entropy (MaxEnt) models or Hidden Markov Models (HMM). However Rule and transformation based taggers often suffice [2–5]. There even exists hybrid models that use probabilistic or stochastic methods in conjunction with rule sets. UCREL CLAWS tagger is an example hybrid model. [**?**] As long as the POS tagging tool utilised performs comparably to those in literature, labouring over the type of POS tagger that is used will not overly affect the results of this study. It might be more useful to consider the subjective merits of any POS tagging libraries that already exist. The role POS tagging plays in this study is to produce a corpus of words that can be filtered by type. When building a model of words associated with entities and deriving topics that describe entities, words such as 'The', 'To' and 'where' won't provide much information about entities within a corpus. Removing them may see an improvement in the quality of any

derived models.

2.3 POS Tagging Techiques and Tools

2.4 Named Entity Recognition

2.4.1 NER Techniques

2.4.2 NER Tools

2.5 Topic Models

2.5.1 Latent Semantic Analysis

2.5.2 Probabilistic Latent Semantic Indexing

2.5.3 Latent Dirichlet Allocation

Bibliography

[1] Web server survey — netcraft. https://news.netcraft.com/archives/category/web-server-survey/. Accessed: 2017-10-16.

[2] E. Brill. A simple rule-based part of speech tagger. In *Proceedings of the Workshop on Speech and Natural Language*, HLT '91, pages 112–116, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[3] E. Brill. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Comput. Linguist.*, 21(4):543–565, Dec. 1995.

[4] D. Cutting, J. Kupiec, J. Pedersen, and P. Sibun. A practical part-of-speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, ANLC '92, pages 133–140, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics.

[5] H. Huang and X. Zhang. Part-of-speech tagger based on maximum entropy model. In *2009 2nd IEEE International Conference on Computer Science and Information Technology*, pages 26–29, Aug. 2009.

[6] G. Leech, R. Garside, and M. Bryant. CLAWS4: The tagging of the british national corpus. In *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, COLING '94, pages 622–628, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.