# 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 [2]. 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

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 [8–10, 14]. There even exists hybrid models that use probabilistic or stochastic methods in conjunction with rule sets. UCREL CLAWS tagger is an example hybrid model. [18] 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

The Natural Language Toolkit (NLTK) by default utilises a Maximum Entropy POS Tagger using the Penn-Treebank tagset. The main benefit to this POS tagger is its ease of use and accessibility. The performance of the MaxEnt tagger used in NLTK was reported score an accuracy of 96.64% on all words, and 85.56% on unknown words. [23] However, in similar implementations, certain tags had error rates of 100% , this was likely a result of the absence of certain features necessary to correctly tag specific parts of speech like 'TO' which occured 14,748 times in a study and was never correctly classified [19].

By comparison, the CLAWS tagger achieved a reported 96% accuracy, though accuracy fell to 82% when text was not preprocessed to filter spelling variants and shakespearean english words. The challenges with POS tagging is the variance of sentence structure, as well as ambiguity that can occur within the english language. Fictional narratives like that of shakespeare frequently contribute to the open classes of words (Adjectives, Nouns, etc), meaning that there is an increased likelihood of unknown or spelling variant words. Whilst modern texts are not likely to be as creative with the english language as shakespeare, it would be naive to assume the CLAWS tagger would achieve the upper bounds for its accuracy. Estimating the performance of the CLAWS tagger on the corpora being used in this investigation makes the performance of both the CLAWS tagger and NLTK's MaxEnt tagger roughly comparable.

In much of the literature, POS tagging techniques were evaluated using corpora oriented around a specific subject or domain, like news, or personal tweets. Using corpora limited to specific domains means that any given POS tagging method might only perform as reported in literature if the same or similar corpus is used. This project will benefit from a POS tagger that performs well on general purpose text, meaning that considering many of the tools used in literature opens a potential point of failure. As POS tagging is one of the first steps performed when processing a corpus, errors could be introduced into the system early on should the POS taggers used in literature proved to be inadequate on the chosen

corpus for this project. Additionally, the effect of literature tending to focus on domain limited corpora means that studies on the quality of general purpose cross-domain POS taggers are somewhat unreported.

## 2.4 Named Entity Recognition

The task of extracting a list of entities from text is a similar task to that of POS tagging. It involves parsing natural language and applying entity labels to entity words. A recent survey of the field summarised that Named Entity Recognition (NER) - the task of identifying entities and labelling them as 'Person', 'Organisation', or 'Location' - as being essential to many tasks of computational linguistics [21]. Similarly to POS tagging, ambiguity prevents NER from having a simple solution. It is not always clear which entity some text may be referring to, especially in situations where an entity is not referred to directly, or when entities are named in unusual ways. [22] Its application in this study is to provide a set of entities from which topic models can be derived from and applied too.

### 2.4.1 NER Techniques

NER can be carried out using a range of different statistical methods. HHMs can be used to classify entities as either a name (person, organisation, or location), time, or numerical quantity. HMM-based Chunk Taggers have a reported accuracy ranging from 87% to 94% depending on the size of the training set [26].

Similarly to POS Tagging, statistical methods based on Maximum Entropy can be applied with similar levels of accuracy as other methods [4,7]. Hybrid approach to NER have been investigated, by utilising HMM, MaxEnt and transformation-based learning, error rates were reduced by as much as 15% when used with the english language [24].

CRFs are a form of statistical model that is particularly useful for applying labels to sequence data, when applied to POS tagging or NER, CRF systems can attain

error rates as low as 5.55% and 15.96% respectively [16, 20].

It is worth noting that the drawbacks that applied to publications regarding POS tagging, also apply to studies on NER. Notably, NER methods typically only concern themselves with labelling words in text, and provide no means of extracting additional information about a recognised named entity. Drawbacks common though many of the techniques revolve around the resolution of which entities are actually the same. Entities within books can be referred to directly, or indirectly, and even be addressed with different names.

### 2.4.2 NER Tools

Many free NER tools exist on the web, including Stanford NER, Illinois NER, OpenCalais NER, and Alias-i LingPipe. In a comparison between the relative performances of these tools to classify entity types (Person, Location, Organisation) the Stanford NER achieved the second highest precision and the highest recall rates. [3]. In separate comparison of NER tagging tools, more mixed results were received for the stanford NER tool. It performed comparably generalised NER tools found in the NLTK and apache OpenNLP toolkits. On specialised datasets, specialised NER taggers tuned to the task at hand predictably outperformed an untuned implementations of the Stanford NER.

Focusing on the Stanford NER [12] an NER tool that utilises conditional random fields (CRF). A Java implementation, the tool exists as part of a larger CoreNLP toolkit created at Stanford University. The self-reported performance of the tool ranges from 92.15% to 85.6% and 92.39% to 85.53% for precision and recall respectively. The upper bounds for possible precision and recall relied on additional processing for handling specific features of text. [1]. Additionally, these results are from tests occuring in 2006. Recent advances in CRFs have been utilised in subsequent versions of the Stanford NER.

## 2.5 Topic Models

Topic models are a way of representing a corpus of text in terms of latent or abstract topics. A topic is defined as the degrees of membership each term in a collection has to a topic. In methods like Latent Dirichlet Allocation, a corpus is expressed as a probability distribution of topics, which in turn are expressed as a probability distribution of words. Describing a specific element of a corpus in terms of topic models requires matching terms in the text to terms in the various topics, after the derivation of a set of models.

Deriving topics from a corpus is commonly done through statistical techniques such as Latent Dirichlet Allocation (LDA), Hierarchical LDA, Latent Semantic Indexing (LSI), or Probabilistic LSI (pLSI). However, many existing methods are in fact built upon the ideas used in LSA, for example, the valuable LDA method arose out of the shortcomings of the comparatively simple LSI [6].

### 2.5.1 Latent Semantic Analysis

Sometimes referred to as LSI, Latent Semantic Analysis (LSA) is a technique that provides vector representations of documents in a corpus. These vector representations allow for quantitative comparisons of different documents. The implementation of the technique is oriented around Single Value Decomposition (SVD), and attempts to model terms in a document as being averages of the document passages which in which it occurs [**?**]. Additionally, for NLP tasks It is possible to view LSA as an extension of the term frequency-inverse document frequency (tf-idf) method, this is due to the method producing a subset of the reduction carried out by tf-idf such that term occurrences with the most variation between documents are retained [17]. The obvious shortcoming of LSA is focused purely on documents at a word level. It bears no notion of topics or themes through which documents can be related. It is possible to apply additional post processing on the output of LDA. Additionally, for this study the method is totally inadequate for the analysis of entities within the documents. The possible use

of the method with regards to the automatic extraction of Latent Entity Models could relate to the extraction of entity-term associations. It has been reported that the method performs well with highly dimensional data, and comparably better than standard vector space models like tf-idf. [15] As such, LSA would likely be useful as nothing more Principal Component Analysis style reduction.

### 2.5.2 Probabilistic Latent Semantic Indexing

Probabilistic LSA (pLSA) is derived as an effort to mathematically formalise the the ideas patented by Deerwester et. al. The formal statistical nature of the method means that the method can be used in conjunction with other models in a more predictable manner. The probabilistic approach changes the meaning of output produced by pLSA, resulting in models which numerically relate terms to some latent variable. When used in tasks of prediction, pLSA models outperformed models produced by LSA. pLSA, reducing the perplexity of any produced models. Notably, there is a greater difference in performance on more general purpose information retrieval tasks, than on restricted corpora. [13]. This is likely due to the effect that sparsity has on the ability of each method. Whilst still worse than pLSA, LSA actually performed better as the sparsity of training data reduced whereas pLSA's performance decreased. This suggests that the mixture models used in pLSA perhaps overfit as the sparsity of data decreases. As this study is focused on written narratives, any corpora used will be diverse and intrinsically sparse. As such the overfitting problem potentially reducing the effectiveness of pLSA might not be so much of an issue. However, whilst it does begin to numerically relate variables, there is still no explicit relation between documents expressed within the models meaning that adding additional documents to the corpus would mean the algorithm needs to be re-run, Additionally, there remains no direct means of relating entities or expressing entities in terms of elements within a corpus.

### 2.5.3 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) uses mixture models to generate a collection of topics which can be used to describe individual documents within a corpus. By representing individual topics as probability distributions of words found within the corpus, additional documents can be expressed as distributions of the topics already derived. This approach overcomes some of the drawbacks of the pLSA method, whilst providing a somewhat human-interpretable representation of the corpus. [6]

For each document win a corpus $D$, LDA produces a mixture model representing proportions of$k$ topics for $w$. Where each word $w_n$ wof length $N$, has a set denoting the probability of $w_n$ belonging to each of $k$ topics. Collections of models could be compared conceptually to weighted groups produced by some clustering method.

The simplified process for generating LDA Models is as follows:
1.$Choose\, N \sim Poisson(\xi)$
2.$Choose\, \theta \sim Dirichlet(\alpha)$
3.$For\, each\, of\, the\, N\, words\, w_n$ :

$a)\, Choose\, a\, topic\, z_n \sim Multinomial(\theta)$

$b)\, Choose\, a\, word\, w_n\, from\, P(w_n|z_n, \beta)$

ENTER PLATE DIAGRAM HERE

In LDA, the lengths of documents are said to be Poisson distributed, and that the set of probabilities generated for each word wnis a random multinomial dirichlet distribution. There is little reasoning for these assumptions. The use of Poisson distributions in literature seems to be a standard distribution for estimating document lengths. However, given that the lengths of documents will be known in this study, a more accurate estimation could be used.

The versatility of LDA is one of its strongest characteristics. The method has seen use in a range of different settings, from filtering web spam, to the semantic annotation of satellite images. As a bayesian statistical process, it is possible to

utilise other methods with relative ease given its modularity.

In web spam filtering, LDA has been combined with tf-idf and expanded to take into consideration links that exist between documents. For spam classification, the use of LDA as a bayesian network performed worse than baseline machine learning methods, however improvements were seen when the method was expanded and combined with other statistical techniques. [5]

For automatic image annotation, additional machine learning and computer vision methods can be applied to images in order transform images into collections of image features. Topics derived from a corpus of these images can be used to annotate new images introduced. Attempts at using LDA in image annotation have seen some success, whilst not all annotations from topics were correct, they were often at least conceptually related. [11, 25]

LDA doesn't relate words semantically. The meanings of words in a topic are not necessarily relate in any way, which is demonstrated in studies using LDA to annotate images. This means that derived topics might not provide any meaningful grouping of words. The intuition is that words in books and documents about a specific subject will somewhat relate to other words in the document. LDA might not produce thematically related topics if the documents used to derive topic distributions are varied in content. For this investigation, the potential for thematically unrelated topics is potentially a major drawback. Entities within narratives might be present throughout a whole series of books. Meaning that they may be associated with a wide range of topics. It is unclear at this stage what effect this may have on the quality of derived topics. It might be necessary to regard the same entity across two books to actually be distinct.

As a bag of words method, each unigram is treated as being unrelated to other words. Which is not true in the case of written narratives. Much of the meaning of words and relations between words are lost when used in 'bag of word' models. The drawback for this study is that relations between terms and entities are required. If topics were directly derived from the documents which entities were extracted from, then there would be no meaningful way of identifying how these

topics apply to entities.

# Chapter 3

# New Ideas

## 3.1 Introduction

# Chapter 4

# Prototyping

## 4.1 Introduction

# Chapter 5

# Results and Evaluation

## 5.1 Introduction

# Chapter 6

# Conclusion

## 6.1 Introduction

# Bibliography


[1] The stanford natural language processing group. https://nlp.stanford.edu/projects/project-ner.shtml. Accessed: 2018-1-10.

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

[3] S. Atdağ and V. Labatut. A comparison of named entity recognition tools applied to biographical texts. Aug. 2013.

[4] O. Bender, F. J. Och, and H. Ney. Maximum entropy models for named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 148–151, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[5] I. Bíró, J. Szabó, and A. A. Benczúr. Latent dirichlet allocation in web spam filtering. In *Proceedings of the 4th International Workshop on Adversarial Information Retrieval on the Web*, AIRWeb '08, pages 29–32, New York, NY, USA, 2008. ACM.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(Jan):993–1022, 2003.

[7] A. Borthwick and R. Grishman. A maximum entropy approach to named entity recognition. 1999.

[8] 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.

[9] 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.

[10] 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.

[11] Y. Feng and M. Lapata. Topic models for image annotation and text illustration. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 831–839, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[12] J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370, 2005.

[13] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[14] 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.

[15] C. A. Kumar and S. Srinivas. On the performance of latent semantic indexing-based information retrieval. *CIT. Journal of Computing and Information Technology*, 17(3):259–264, Oct. 2004.

[16] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289. Morgan Kaufmann Publishers Inc., June 2001.

[17] T. K. Landauer, P. W. Foltz, and D. Laham. An introduction to latent semantic analysis. *Discourse Process.*, 25(2-3):259–284, Jan. 1998.

[18] 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.

[19] G. Malecha and I. Smith. Maximum entropy part-of-speech tagging in NLTK. *unpublished course-related report*, 2010.

[20] A. McCallum and W. Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003 - Volume 4*, CONLL '03, pages 188–191, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[21] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, 30(1):3–26, Jan. 2007.

[22] L. Ratinov and D. Roth. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning*, CoNLL '09, pages 147–155, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[23] A. Ratnaparkhi. A maximum entropy model for part-of-speech tagging. *of the conference on empirical methods in ...*, 1996.

[24] E. Tjong Kim and F. De Meulder. Introduction to the CoNLL-2003 shared task: Language-Independent named entity recognition. June 2003.

[25] T. Zhang, Z.-M. Lu, K. L. Chan, and Z. Li. Automatic image annotation and retrieval using the latent dirichlet allocation model. *IJCSES International Journal of Computer Sciences and Engineering Systems*, 5(1), 2011.

[26] G. Zhou and J. Su. Named entity recognition using an HMM-based chunk tagger. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 473–480, Stroudsburg, PA, USA, 2002. Association for Computational Linguistics.