BIROn - Birkbeck Institutional Research Online

Harris, Martyn and Levene, Mark and Zhang, Dell and Levene, D. (2018) Finding parallel passages in cultural heritage archives. *ACM Journal on Computing and Cultural Heritage* 11 (3), ISSN 1556-4673.

Downloaded from: <https://eprints.bbk.ac.uk/id/eprint/21385/>

Usage Guidelines:

Please refer to usage guidelines at <https://eprints.bbk.ac.uk/policies.html>
contact lib-eprints@bbk.ac.uk.

or alternatively

Finding Parallel Passages in Cultural Heritage Archives

MARTYN HARRIS, MARK LEVENE, and DELL ZHANG, Birkbeck, University of London
DAN LEVENE, Southampton University

It is of great interest to researchers and scholars in many disciplines (particularly those working on cultural heritage projects) to study parallel passages (i.e., identical or similar pieces of text describing the same thing) in digital text archives. Although there exist a few software tools for this purpose, they are restricted to a specific domain (e.g., the Bible) or a specific language (e.g., Hebrew). In this paper, we present in detail how we build a digital infrastructure that can facilitate the search and discovery of parallel passages for any domain in any language. It is at the core of our Samtla (Search And Mining Tools with Linguistic Analysis) system designed in collaboration with historians and linguists. The system has already been used to support research on five large text corpora that span a number of different domains and languages. The key to such a domain-independent and language-independent digital infrastructure is a novel combination of a character-based n -gram language model, space-optimised suffix tree, generalised edit distance. A comprehensive evaluation through crowd-sourcing shows that the effectiveness of our system's search functionality is on par with the human-level performance.

CCS Concepts: • **Information systems** → **Information retrieval**; **Information retrieval query processing**; **Language models**; **Evaluation of retrieval results**; • **Applied computing** → **Document searching**;

Additional Key Words and Phrases: Digital Archives, Information Retrieval, Statistical Language Models, Suffixes Trees

ACM Reference format:

Martyn Harris, Mark Levene, Dell Zhang, and Dan Levene. 2018. Finding Parallel Passages in Cultural Heritage Archives. *ACM J. Comput. Cult. Herit.* 11, 1, Article 1 (February 2018), 26 pages.
<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

1 INTRODUCTION

Digitised textual representations of original historic documents are being made available at unprecedented speed, thanks to many culture heritage projects [47]. As the wealth of digital text archives rapidly increases, it becomes more important to provide text analytic tools to assist the research of big data represented by large scale archives [24]. Tools and systems have been developed to support humanities researchers in analysing primary source material, however they have not generally been widely adopted, meaning that researchers are not making full use of digital archives as a key resource for discovering re-occurring cultural contexts that can help to address new research questions, and revisit old ones through a much larger body of evidence provided by these digital repositories.

One text analysis problem that is of great interest to researchers, particularly those analysing the content of digital archives, is to find *parallel passages* — text segments describing the same concept (entity or event etc.) over

We would like to thank the anonymous reviewers for their time and useful comments, which helped to improve the original draft of this paper. This work was partly supported by the Wellcome Trust Institutional Strategic Support Fund (ISSF). Authors' addresses: Martyn Harris, Mark Levene, Dell Zhang. Department of Computer Science and Information Systems, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK; Dan Levene, Building 65, Faculty of Humanities, University of Southampton, Avenue Campus, Highfield, Southampton SO17 1BF, UK. .

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

© 2018 Copyright held by the owner/author(s).

XXXX-XXXX/2018/2-ART1

<https://doi.org/10.1145/nnnnnnnn.nnnnnnnn>

Table 1. An example “parallel passage”, adapted from [14].

<i>2 Kings, Chapter 16</i>	<i>Isaiah, Chapter 7</i>
Then	In the days of Ahaz son of Jotham son of Uzziah, king of Judah
came up	came up
King Rezin of Aram and	King Rezin of Aram and
King Pekah son of Remaliah of Israel	King Pekah son of Remaliah of Israel
to wage war on Jerusalem;	to Jerusalem to attack it,
they besieged Ahaz	
but could not	but could not
prevail over him.	mount an attack against it.

large corpora. Parallel passages are semantically similar and could exhibit identical wording, but quite often they exhibit some small variation in structure, or vocabulary choice. The differences are due to the normal rephrasing of the text within the same context, but may also arise from the use of reported speech, a change in authorship, dialect, the natural evolution of language over time, and errors introduced by optical character recognition (OCR). The term parallel passage probably originated from Christian theology, where the comparison of parallel passages in the context of the Bible is a major area of Biblical scholarship (see [52]). The technique involves comparing corresponding passages located across more than one text, by laying out the texts side-by-side. The Bible often describes the same event from different perspective across different canonical books, which can yield a more complete picture of the event than a single passage, or point of view on the subject. The example presented in Table 1 is from the *King James Bible*, and illustrates two “parallel passages” that could be regarded as highly similar by researchers of the Bible. These texts discuss the same event, and it has been proposed that *Isaiah, Chapter 7* was derived from the text of *2 Kings, Chapter 16* [14]. The similarity between these two texts is not easily identifiable with current tools developed for search and mining of digital archives, due to the variability in the structure and choice of vocabulary, where *2 Kings, Chapter 16* adopts the phrase “wage war” over the word “attack” in *Isaiah, Chapter 7*, and the latter text is also more specific about the time of the event, as described in the introductory section of the text.

It is not difficult to imagine the existence and usefulness of parallel passages in other domains. For example, in newspaper archives, it is common to see the same news event being covered by many news reports from different sources and at different time points: those news reports would contain a number of parallel passages and their mutual corroboration could help readers gain a bigger picture or balanced view of an event. In this paper, the concept of parallel passage is defined in the general sense. Roughly speaking, one can regard parallel passages as a variable length structural text pattern. Although there exist a few software tools for finding parallel passages, they tend to be quite limited in terms of functionality, and more importantly, they are often restricted to a specific domain, or language (see Section 2). In this paper, we present a digital infrastructure that greatly facilitates the finding of parallel passages in any domain or language. Here the notion of “finding” is divided into *search* (i.e., locating exact and approximate parallel passages with respect to a given text segment, across multiple documents), and the other refers to *discovery* (i.e., identifying potentially related text segments). The *domain-independent* and *language-independent* digital infrastructure introduced in this paper is at the core of the Samtla (Search And Mining Tools for Language Archives) system¹ [21, 22] which has been designed in collaboration with historians, and linguists. Samtla is a web application, based on the *a Model-View-Controller (MVC)* design pattern [33], and

¹<http://www.samtla.com/>

Table 2. The current archives supported by Samtla.

Archive	Documents	Size	Language(s)
Aramaic Magic Text from Late Antiquity (Southampton University)	539	830.5KB	Jewish Aramaic, Syriac, Mandaic.
UK Medical Heritage Library (Wellcome Trust)	75,973	35.1GB	English (Middle, Modern), French, Spanish, Italian, Russian.
Financial Times Historic Archive (British Library)	70,640	230MB	English.
Microsoft Corpus of Scanned Books (British Library)	65,236	241.5GB	English, French, Spanish, Italian, Hungarian, Russian, Malagasey.
Works of Giorgio Vasari Dept. Art History, (Birkbeck University)	314	5.1MB	English, Italian.
English Bibles (1500 - 2000) (Archive.org)	185	947.2KB	English (Early Modern to Modern).

developed in Python² using the Django³ web framework. The *Samtla* system operates with a single code-base making the whole system data-driven. All digital text archive data, and supporting index are stored on a central server, and the application is accessed through a modern web browser. The Samtla system supports the research of several text corpora that span a number of different domains and languages, and vary in scale, which we summarise in Table 2.

The *Aramaic Magic Bowls and Amulets from Late Antiquity*⁴ (6th to 8th CE) are the focus of research of a team of historians from Israel and the UK led by the University of Southampton. The texts were written in ink on clay bowls using a number of related dialects including Aramaic, Mandaic, and Syriac. Samtla was originally developed in response to a practical need for flexible search and comparison tools by this team of researchers to assist them with locating and comparing formulaic parallel passages that can provide an insight into the evolution and transmission of liturgical forms over the centuries. The *UK Medical Heritage Library*⁵ is a collection of more than 70,000 digitised texts from the history of medicine in the 19th century. It was curated by the Wellcome Library along with nine partner institutions. The documents in this archive span a broad range of topics, including medical articles, health reports, diet and nutrition, and historical documents relating to medical practices (e.g. phrenology). The archive also contains documents in several languages, including English⁶, French, German, Spanish, Italian, and Russian. The *Financial Times (FT) historic newspaper archive* (covering the year 1888, 1939, 1966, and 1991) and published in English, formed the basis of a pilot study organised by the British Library and the Financial Times. The Microsoft - British Library collection of 68,000 digitised books from the world's libraries, contains approximately eight different languages, and a variety of literary genres, and periods from the 15th century to the mid-19th century. The *British Library Microsoft archive* contains documents with very poor quality OCR, or no textual content at all when the documents represent images of plates, figures, and illustrations. The works of Giorgio Vasari, known as the *Lives of the Most Excellent Painters, Sculptors, and Architects* (1511–1574),

²<https://www.python.org/>

³<https://www.djangoproject.com/>

⁴<http://www.southampton.ac.uk/vmba/>

⁵<http://wellcomelibrary.org/collections/digital-collections/uk-medical-heritage-library/>

⁶Among the English documents, the UK Medical Heritage Library explicitly lists four of them in Middle English.

contains documents in the original Italian as well as a corresponding English translation and is used for research and teaching by the Vasari Research Centre⁷. The *Bible*, in English, is used for demonstration and evaluation purposes as many people are familiar with the content of the Bible.

The Samtla system operates with a single uniform and consistent framework, thanks to its domain-independent and language-independent design. The only corpus specific component of the infrastructure is a wrapper for parsing the documents and the associated metadata into the system according to a number of popular formats including, plain text, HTML, XML, TEI, and PDF. Samtla was developed due to a lack of tools sufficient for searching and comparing text patterns whose internal structure may be quite different despite similarities in the semantics of the passages. We believe that such a system could be utilised to advance digital humanities projects such as Ryan Cordell’s “The Celestial Railroad” project [11] which looks at the reprinting and recompositing of texts across a large corpus. The flexibility of Samtla to search across a combination of synonyms, other spellings, and structure would be especially valuable for this sort of work.

Although the full Samtla system provides a comprehensive suite of “search and research” tools for text retrieval and mining (discussed in a companion paper to follow), this paper discusses the components of the digital infrastructure used for the search and discovery of parallel passages. The key to make such a digital infrastructure is a novel combination of *character-based n-gram language model*, *space-optimised suffix tree*, and *generalised edit distance*. A comprehensive evaluation through crowd-sourcing shows that the effectiveness of our system’s search functionality is on par with the human-level performance.

The paper is organised as follows. In Section 2, we review the related work. In Section 3, we describe the underlying model of our digital infrastructure. In Section 4, we present our methods for the search and discovery of parallel passages. In Section 5, we present several case studies demonstrating the flexibility of the approach to a range of digital archives. In Section 6, we present the results of a formal evaluation that demonstrates the effectiveness of our system. We conclude the paper with a summary of the work in Section 7.

2 RELATED WORK

Humanities researchers are discovering how digital tools can become part of their methodology. There are several barriers to computer-assisted forms of analysis in the humanities, some are related to the incompatibility of the tools with the research topic, methodology, or digital archive of interest [18, 53], due to the fact that they are often developed as part of specific research projects, and are not designed to be inter-operable across other data standards, formats, domains, or languages. Further barriers relate to the way the tools are implemented, where many tend to be domain and language-specific due to the word-level model commonly adopted as the document and user query representation. A word-level model requires the texts to be preprocessed through language-specific segmentation and normalisation of the text in order to identify all occurrences of the same term regardless of the syntax of the language. However, the word-level model is not a suitable representation when the language contains no explicit word-delimiter, when the morphology of the language is more complex than that of the commonly studied languages, or when applied to historic documents containing non-standard spelling and formatting. It is apparent that the current word-level representation will become difficult to sustain in future due to the increasing diversity and large-scale nature of digital archives.

A review of digital tools available to researchers highlights some notable examples. The Bar-Ilan *Responsa* project⁸, established in 1963, is one of the earliest examples of humanities researchers adopting the use of computing techniques for the retrieval and analysis of Hebrew texts. The corpus spans approximately 3000 years, and includes the Mishnah, Talmud, Torah, and the Bible in Aramaic [10]. As another example, the *1641 Depositions*

⁷<http://www.bbk.ac.uk/arts/research/vasari-research-centre/>

⁸<http://www.biu.ac.il/JH/Responsa/>

project⁹ at Trinity College Library (Dublin) uses IBM's LanguageWare to analyse 31 books containing 19,010 pages of witness accounts reporting theft, vandalism, murder, and land taking during the conflicts between Catholics and Protestants in the 17th century Ireland. However, these large-scale general-purpose digital humanities projects do not usually have dedicated powerful tools for discovering parallel passages in digital archives through search. There are, however, a few notable examples of software tools for analysing parallel passages, including the *Logos Bible*¹⁰ system and the *BibleWorks*¹¹ system, which allow users to display and compare parallel passages in the Bible as inter-linear text. The *SHEBANQ*¹² toolkit is used for detecting parallel passages for the study of the Hebrew Bible, whereas the *Chinese Text Project*¹³ contains an integrated database of hundreds of thousands of parallel passages in early Chinese texts, the majority of which have been identified by custom-made software tools. *Voyant Tools*¹⁴ is a browser-based suite of text-analysis tools that enable researchers to browse and analyse a corpus of documents [31, 46]. The system provides tools for generating concordances through Key Word In Context (KWIC) [39] and word clouds for identifying the most frequent terms in a document and the corpus etc. In addition, the tool provides data-analytics in the form of statistics constructed from the word frequency distributions, according to specific sections of the document.

The current set of software tools are still of quite limited functionality (e.g., dependence on manual annotations), or tied to a specific archive or language. Furthermore, it is also unclear what approaches these software tools adopt for finding parallel passages, as there does not seem to be any research literature on this topic despite its practical importance, as far as we know. Standard *Information Retrieval* (IR) [38] systems, including open source search engines like *Lucene*¹⁵ and commercial search engines [34] like *Google*¹⁶, are designed to process *keyword queries*, but not to find parallel passages. Although many of them support *phrase search* for locating exact matches for short text segments, they cannot identify partial matches necessary for solving the problem of finding parallel passages, where constituents of the phrase may be absent, or spelt differently. For example, if a user searches the query "As *the* Lord commanded . . ." in the Bible, we would also expect to locate partial matches like "As *thy* Lord commanded . . ." despite the surface difference between "*the*" and "*thy*". There are also the so-called *tolerant search* techniques, but they are for the purpose of generating spelling corrections or handling wildcard queries, which is a different problem to locating parallel passages.

It might be worth to point out that the concept of "parallel passages" in this paper is connected to the so-called "parallel corpora" in machine translation [9, 32, 36, 54]. Their discovery or extraction would require very different techniques, though: the former can exploit the surface string similarity between parallel texts which are in the same language (as we do in this paper), but for the latter there is apparently none or little surface string similarity between parallel texts due to the fact that they are in completely different languages so it has to rely on word-based alignment rather than character-based n-gram statistical language models.

3 MODEL

The software architecture of Samtla is shown in Figure 1.

The digital infrastructure supporting Samtla is based on character *n*-gram *Statistical Language Models* (SLM) [56] (see Section 3.1). We store the characters of the documents in a character-based *suffix tree* [20] data structure to produce an index of the *n*-gram sequences for each document (see Section 3.4). The combination of a character-level

⁹<http://www.1641.tcd.ie/project.php>

¹⁰<https://www.logos.com/media/tour/ParallelPassages.htm>

¹¹http://www.bibleworks.com/classroom/1_10/

¹²<https://shebanq.ancient-data.org/>

¹³<http://ctext.org/tools/parallel-passages>

¹⁴<http://voyant-tools.org/>

¹⁵<https://lucene.apache.org/>

¹⁶<https://www.google.com/>

Fig. 1. The Samtla architecture.

representation makes query processing and further analysis very flexible, and the *SLM* represents a probabilistic approach that is well motivated, and understood by many researchers across disciplines, which makes the underlying model transparent in terms of how the tools interface with the data in order to help researchers adopt it as an integral part of their research methodology [18, 53].

3.1 Statistical Language Models

An SLM is a mathematical model representing the probabilistic distribution of words or sequences of characters found in natural language represented by text corpora [37, 48, 56], and provides a consistent methodology for retrieving and ranking search results according to the underlying principles and structure of natural language, which is often domain specific. A character-based n -gram SLM, rather than the more conventional word-based model, enables the system to handle multilingual corpora with very little pre-processing of the documents. For instance, languages such as Hebrew, and Russian attach affixes to a root morpheme to identify syntactic relationships. Word-based retrieval models require language-dependent stemming and part-of-speech tagging algorithms, whereas a character-based n -gram model resolves some of these issues, and have been shown to outperform raw word-based models when the language is morphologically complex [29, 40, 60]. The character-based model can also be applied to corpora containing documents written in different languages. For example, the British Library Microsoft Corpus and the Wellcome Trust UK Medical Heritage Library both hold documents in a range of morphologically different languages including English, French, Hungarian, and Russian.

Operationally, when the user submits a query, a list of documents is ranked according to how relevant the document is to the query. The notion of relevance refers to the users expectation of which documents should be present at the top of the ranked list [56]. We take the view that the more probable a document in the SLM sense, the more relevant it is to the user, which avoids much of the philosophical debate on the notion of “relevance” [43].

We generate SLMs from the whole document collection, called the *collection model* C , and over each individual document, known as the *document model* D , which together represent the basic language model [57].

Metadata search has been shown to be a particularly effective form of support for search, since authors may not make explicit mention of the subject matter of the document in the body of the text [12], and it also supports the discovery of documents across different media, e.g. images, video, and audio. We extend the basic language model through an additional *metadata model* B representing a SLM constructed from the metadata record for each document stored in an additional suffix tree data structure. The only difference between the metadata model B and the collection model C , is the further storage of the metadata field column label used to construct a metadata search filter tool (discussed in Section 5).

Each SLM is generated from the character-level n -grams of the document, where n will vary from one to some pre-determined maximum, which are then reduced to lower-order n -grams by removing a character from the sequence at a time. Throughout the remaining sections of the paper, we denote a generic SLM by M .

3.2 Query processing

If we let $q = c_1, c_2, \dots, c_m$ denote a sequence of text consisting of m characters, then the probability of q being generated by the language model M , denoted by $P(q|M)$ or $P_M(q)$, can be calculated using the chain rule as a product of conditional probabilities:

$$P_M(q) = P_M(c_1, c_2, \dots, c_m) = \prod_{i=1}^m P_M(c_i | c_1, \dots, c_{i-1}). \quad (1)$$

Each conditional probability, $P_M(c_i | c_1, \dots, c_{i-1})$, on the right-hand side of (1), with $1 \leq i \leq m$, is approximated by the maximum likelihood estimator (MLE) [38]:

$$MLE_M(c_i | c_{i-n+1}, \dots, c_{i-1}) = \frac{\#(c_{i-n+1}, \dots, c_i)}{\#(c_{i-n+1}, \dots, c_{i-1})}, \quad (2)$$

where the $\#$ symbol indicates the raw count of the sequence in the model M . For any character sequence we only make use of its n character history (or less than n for shorter sequences) to approximate the conditional probabilities in (1) according to the n -order Markov rule [41] (We also take $MLE_M(c_1 | c_0)$ to be $MLE_M(c_1)$.) The probabilities are assigned to each node in the suffix tree through a traversal of the data structure starting at the root node, where we calculate the conditional probability for each node by dividing its count by that of the parent node, as defined in (2).

3.3 Smoothing the language model

Smoothing is an important component of a SLM [8, 59]. When the count of a n -gram of the query is zero the query model probability, $P_M(q)$ in (5), is also evaluated as zero even if some n -grams of the query are located. This may be caused by spelling mistakes or typographical errors in the query, or the document due to poor OCR. To overcome this issue we smooth the MLE probability (2) of a n -gram with lower-order n -grams that have more reliable counts through interpolation, using a weighted term. The weighted term defines the contribution of the probability for each order of n -gram, k , where k varies from a zero order, 0-gram model, when $k = n + 1$, to an n -gram model, when $k = 1$. Each weight, represented by λ_k , defines the amount of interpolation, with lower-order models contributing less to the final probability. Consequently, the approximation of the conditional probabilities on the right-hand side of (1) is given by the interpolation,

$$\hat{P}_M(c_i | c_1, \dots, c_{i-1}) \approx \sum_{k=1}^{n+1} \lambda_k MLE_M(c_i | c_{i-n+k}, \dots, c_{i-1}), \quad (3)$$

Fig. 2. A suffix tree for the text string “xabxa”.

where we use $\hat{P}$ to indicate that we are approximating P . The weighted term for each k is given by

$$\lambda_k = \frac{n + 2 - k}{(n + 1)(n + 2)/2}, \quad (4)$$

where n is the order of n -gram, which is composed by interpolating the n th order model with lower order ones. When $k = n + 1$, then $MLE_M(c_i|c_{i+1})$ is taken to be the 0-gram model, $\frac{1}{|V|}$, where V is the finite alphabet of the language (for English $|V|=27$, representing the letters of the Roman alphabet, plus the whitespace character).

3.4 Storing the language model

Samtla’s search capability is supported by a space optimised character-based suffix tree, which provides fast and flexible retrieval [20]. We create a k -truncated suffix tree [50] by limiting the depth of the tree to a maximum of k nodes, which reduces the memory consumption. We have found that $k = 15$ works well for the languages we have experimented with, as the average word length has tended to be no more than 15 characters. Given a text string, the resulting generalised suffix tree represents a compressed “trie” data structure with the suffixes of the string as their keys and positions in the string as their values, and the leaf nodes store the document ID and start position of the character string.

4 SEARCHING PARALLEL PASSAGES

When a query is submitted, we compute the probability that the query was generated by the model, M , where M is either D , C or B . The documents are ranked according to the computed probabilities and the top scoring documents are returned to the user. The search begins at the root node of the suffix tree where we descend the tree along a unique path by comparing characters of the query with the label of each node visited; this is performed in linear time according to the size of the query. When the characters of the query are exhausted or a mismatch occurs, the sub-tree rooted at the last-matched node is traversed with a breadth-first traversal, and all leaf nodes are collected to form an index containing every document ID and start position for full and partial matches to the query. Figure 3 shows an example of the Samtla system search.

Both the collection model C and metadata model B are traversed simultaneously, and we obtain an index from the suffix tree (see Section 3.4). Some documents may only retrieve partial matches for the query, which may still be of interest to the researcher. These are ranked lower down in the search results under full query matches to reflect this. To estimate the relevance of a document D with respect to the given query q we calculate the conditional probability of a document D generating the query q based on the distribution of n -grams in the document model D . According to the Bayes theorem [19], we have

$$P(D|q) = \frac{P(q|D)P(D)}{P(q)}. \quad (5)$$

Fig. 3. The Samtla search interface.

The right-hand side of (5) consists of the query model $P(q|D) = P_D(q)$ multiplied by $P(D)$, the *prior* probability of the document D which is often assumed to be uniform (i.e. the same for all documents) and is ignored for the purpose of ranking. The probability of the query $P(q)$ is also the same for all documents. When the prior is uniform, we rank the documents according to the query model $P_D(q)$ for the document. Next, the document model is interpolated with the collection model probabilities to obtain a better global estimate of the probability of the query. We adopt the Jelinek-Mercer smoothing method, which involves a linear interpolation of the document model D , collection model C , and metadata model B , through a coefficient represented by λ to control the influence of each model on the query score. [58]. The final smoothed query score for each document is obtained as follows:

$$P_D(q) \approx \lambda_1 \hat{P}_D(q) + \lambda_2 \hat{P}_C(q) + \lambda_3 \hat{P}_B(q), \quad (6)$$

where we replace $\hat{P}$ with the term P , and λ_i is the weighted term that defines the contribution according to each of the SLMs represented by the document model D with $\lambda_1 = 0.5$, collection model C at $\lambda_2 = 0.4$, and the additional metadata model B through $\lambda_3 = 0.1$. The weights were derived empirically with feedback from our research groups and the weighting scheme ensures that matches for the query in the document text are given more weight as they provide a better description of the topic of a document. Our researchers submit long verbose queries to locate parallel passages, which impact on retrieval performance [25] due to the presence of uninformative query terms representing the syntax of the language. This is because even after the initial interpolation in (2), the probability of a n -gram given the document model may be relatively low due to a low number of occurrences,

however the interpolation of the collection model provides a better estimate based on more information. The λ parameter can also be tuned by experimentation, but in general, long verbose queries require more smoothing with a higher setting for λ [58]. The Jelinek-Mercer smoothing method was adopted as it has been shown to be particularly effective for a range of different query types [58, 59], and is well-suited when smoothing n -gram distributions generated by long verbose queries.

When n -grams of the query are not found in the document or metadata record, we compensate by backing off to a lower order n -gram to approximate the missing probability for the n -gram of the query. A weighted term adjusts the contribution of the missing n -gram to match the appropriate lower order n -gram model. If a lower order n -gram is not located, then we store the intermediary result of μ , and continue to backoff until a lower order n -gram is found, or we backoff to the 0-gram model. More formally, we have

$$\hat{P}_M(c_i | c_1, \dots, c_{i-1}) \approx \sum_{k=1}^{n+1} \begin{cases} \mu_n P_M(c_i | c_{i-n+1}), & \text{if } P_M(c_i | c_{i-n+1}) > 0 \\ \mu_n, & \text{otherwise} \end{cases} \quad (7)$$

where n is the length of the lower order n -gram we are backing off to, and $\mu_n = \frac{n+1}{n+2}$ is the weighted term defining the contribution of the lower order n -gram probability to the approximation of the full query, if it exists, otherwise, we store the result of μ_n , and continue to back off.

The approach introduced in this section provides a lot of flexibility with respect to extending the basic language model with any number of separate language models using a weighted term to control the contribution of each model. The flexibility of the approach is demonstrated by the range of digital archives that are supported, for example, the *British Library Microsoft archive* contains only a metadata model B , and so the lack of the document model D and collection model C components are compensated for in the query model through the backoff process, where the probability for the n -gram of the query given a missing SLM is evaluated on the basis of the 0-gram model (representing $\frac{1}{|V|}$, where $|V|$ is the finite alphabet of the language).

To summarise, smoothing is an important aspect of Statistical Language Models to produce a more accurate model of the n -grams than that provided by the MLE alone, which maximises the likelihood function based on the observed data, but does not take in to account unseen data such as missing n -grams in the document [45]. The interpolation method incorporates information from lower order n -grams, which have more stable counts for approximating the probability for each given n -gram. The interpolation between the collection model C , and the document model D , compensates for the non-descriptive terms of the query, whilst the backoff process handles missing terms of the query as a result of poor query formulation, or spelling errors in the document text.

4.1 Search result snippet generation

Snippets are one of the most common approaches adopted [23] for displaying matches to the query in the document content. Each snippet represents a summary of the document generated from the index returned by the retrieval model. Snippets are classified as either static or dynamic, depending on their application in a system. Static snippets return the same summary of the document each time they are generated, whereas dynamic snippets are created in conjunction with the scoring of queries, with each occurrence of the query highlighted in the document. The snippets are composed of one or more summaries containing all or part of the query terms, with additional context provided to the left and right of the query matches. The challenge with snippet generation is to capture coherent portions of text as users tend to prefer summaries that are easy to interpret [38], and so the aim is to produce snippets that are informative of the topic described by the query, highly readable to facilitate the user in choosing the most relevant document, and concise to make the best use of the space in the user interface when presenting a long list of search results [38].

The *Samtla* system produces dynamic snippet windows from the index returned by the suffix tree at query time. The index provides the means for extracting the matching portions of the query from the document text,

Table 3. Snippets generated for a single document matching the query “Hast thou appealed unto Caesar?”, submitted to the *King James Bible*.

Rank	Score	Snippet text
1	0.648	...answered, Hast thou appealed unto Caesar ? unto Caesar shalt thou go. And after certain days king Agrippa unto Caesar ea...
2	0.072	...he went down unto Caesar ea; and the next day...
3	0.072	...I appeal unto Caesar . Then Festus, when he had co...
4	0.054	...I commanded him to be kept till I might send him to Caesar . Then Agrippa said unto...
5	0.042	...Paul should be kept at Caesar ea, and that he himself would depart...

using the start and end positions of each matching n -gram of the query. The extracted sequences are expanded to the left and right to provide the user with some context to the query. The length of the context is tunable and defined by a parameter w reflecting the maximum length of the context in characters, where $w = 100$, for our purposes. In future versions this could be provided as a user setting. The snippets are scored by calculating the length of the matching n -grams found in the snippet, which is interpolated with the total count of all n -grams found in the snippet. The algorithm assigns more weight to snippets containing all of the query n -grams to ensure the snippets are ranked in such a way that the top snippets will contain full matches to the query, before presenting partial matches. The score for each snippet candidate is defined as

$$\text{SnippetScore} = \delta^\alpha \mu^{(1-\alpha)}, \quad (8)$$

where δ represents the cardinality of the set of n -grams in the snippet, and μ is the count of all n -grams (including repetition). The two terms are then interpolated with a weight defined as $\alpha = 0.9$. The high setting for α ensures that the snippets are biased towards those that contain full matches for the query. The snippets are sorted in descending order according to their respective score, and the top-three are selected as a preview for the document. As an example, Table 3 displays the potential snippets generated for the query “Hast thou appealed unto Caesar?” together with the score for each snippet given by (8). When displaying the search results, the documents are organised according to exact and partial matches for the query by partitioning them in to bins according to the maximum length of the matched query. The documents in each bin are then sorted by probability inferred from the *SLM*, which ensures that users will always be presented with full matches for the query at the top of the search results, before any partial matches are presented. Partial matches do not encompass the full query, but may still be of interest to the user, or may aid the user in reformulating the terms of their query, in order to refine the search results. A further component of the snippet window is the metadata snippet text. The matches for the query given the metadata model B , are displayed at the bottom of each document snippet with the metadata field and the query highlighted in the corresponding record value.

4.2 Query Recommendation

The related queries tool in *Samtla*, addresses many of the language-specific issues, such as differences deriving from the syntax of the language e.g. the encoding of the past tense of the verb or attaching affixes to nouns to describe plurality. Furthermore, an archive can contain documents spanning several time periods. Under this context the system needs to compensate for language change by making a distinction between old forms of the language and their modern day equivalents. There exist many studies on the standardisation or normalisation of spelling for historical texts like those representing earlier forms of English [5, 6]. To illustrate, the editions of the Bible over the centuries reflect updates to the language as it was recorded at the time of publication. As a result

Table 4. Related queries generated for the character-sequence “lord”.

Position	Deletion	Substitution	Insertion
1	ord	?ord	?lord
2	lrd	l?rd	l?ord
3	lod	lo?d	lo?rd
4	lor	lor?	lor?d
5			lord?

we find alternative spellings, such as “Lord” versus “Lorde”, and “Lords” versus “Lordes”. These language-specific differences are recorded in the suffix tree as a series of unique paths rooted at the subtree for the sequence “Lord”, and we can leverage this structure for generating related queries. We make a distinction between *Type I related queries*, which are extracted automatically from the suffix tree component based on a *generalised edit distance* approach, and *Type II related queries*, which are defined through a small set of rules provided by the researchers, before being retrieved from the suffix tree.

4.2.1 Type I related queries. Related queries classified as *Type I* are located by traversing unique paths in the suffix tree through permutations on the order and presence of characters given the full-sequence represented by the query. The related queries are generated through an online process at query time, using a method similar to the Levenshtein edit distance algorithm, which describes the minimum number of operations required to convert one string into another [20, 35]. The method adopted here, produces a series of alternative queries through deletion, substitution, and insertion of the characters of the original query. If we let q' represent the related query, where n is the length of the original query q , and $i \in 1, 2, \dots, n$, then the related queries are generated through the following string permutation methods, where “?” represents the wild-card character, which is a unique character not indexed in the suffix tree collection model C .

- (1) Deletion: $q' = c_1, \dots, c_{i-1}, c_{i+1}, \dots, c_n$.
- (2) Substitution: $q' = c_1, \dots, c_{i-1}, ?, c_{i+1}, \dots, c_n$.
- (3) Insertion: $q' = c_1, \dots, c_{i-1}, ?, c_i, \dots, c_n$.

Table 4, presents an example, where the original query is “lord”. Note that deletion does not require a wild-card character, as such, since we simply remove a character from the string where the wild-card character would appear. The related queries are then submitted to the suffix tree at the same time as the original query. As the wild-card character is not indexed by the suffix tree there will be a guaranteed mismatch at that position in the query. When this occurs we execute each of the above functions, which traverse the suffix tree from the node where the mismatch occurred. Insertion and Substitution are achieved by replacing the wild-card character with the node label of each child rooted at the last matched node (its parent node). We then attempt to match the remainder of the query along a unique path to a leaf node. If there is a match for the remainder of the related query, then we extract the smoothed probability from the last matched node. The extracted queries are then ranked to produce a list of related queries, where those with the highest probability are considered to be the most “related” to the user’s original search. To illustrate, given the sequence “lord?”, the plural form of the word “lords”, and “lorde” are ranked as the top-two related queries, since they represent common permutations according to the morphological rules of old and modern forms of English.

Several examples of the output generated by the *Type I related queries* tool are listed below, divided according to each case study. Each example, includes a description of the type of related query (in brackets), where *syntactic* refers to differences in the grammar of the language, *orthographic* differences represent different lexical forms or an older spelling variant, and *spelling error* refers to an OCR, or transcription error recorded in the document text.

Some related queries represent different languages from that specified by the query, which are identified by the name of the language next to the appropriate related query.

- **UK Medical Heritage Library.**
 - disease → diseases (syntactic), diseased (syntactic), diseaso (spelling error).
 - heart → hart (syntactic), heart (syntactic), hea3t (spelling error).
 - midwife → midwive (syntactic).
 - diarrhoea → diarrhoeal (syntactic).
 - tuberculosis → tuberculosis (syntactic)
- **FT newspaper archive.**
 - Journalist → Journalists (syntactic), Journalism (syntactic).
 - American → Americans (syntactic), Americano (Spanish), Amerlcan (spelling error).
 - Vietnam → Vietnan, Vetnam, Vtetnam, Vienam (spelling errors).
- **British Library Microsoft archive.**
 - India → Indian (syntactic).
 - Poets → poems (syntactic), ports (orthographic).
 - Russia → Russian (syntactic), Prussia (orthographic).
- **Aramaic Magic Bowl archive.**
 - לִילִיחָא “female demon” → לִילִיא “male demon/night” (syntactic), וּלִילִיחָא “and female demon” (syntactic)
 - וְלֹא חִילוּט “that she may not curse” [28] → וְלֹא יִלוּט “that he it may not curse” (syntactic), וְלֹא חִילוּט “that she may not curse” (orthographic) [1].
- **King James Bible.**
 - Lord → Lords (syntactic), Lods (spelling error), Word (orthographic).
 - His → hys (orthographic).
 - Jacob → Iacob (orthographic).
- **Giorgio Vasari archive.**
 - Andrea → Andreas (orthographic), Andrew (English).
 - Sculptor → Sculptors (syntactic), Scultor (Italian).
 - Veneziano → veniziani (orthographic), veneziano (English).

4.2.2 Type II related queries. For some languages it can be challenging to describe all possible string permutations using this approach alone, particularly for documents representing historic corpora. A further component of the related query tool enables researchers to define a small set of rules, or character-mappings, that replace certain characters in the query with a corresponding character-sequence to describe specific differences in morphology, dialect, or spelling. For example, in the *Aramaic Magic Bowl archive*, the rules describe phonological differences between dialects, such as long vowels, (e.g. *i* → *ii*), and the mapping of characters from one writing system to another (e.g. Aramaic script to the Syriac script). The *Bible* archive, on the other hand, contains a selection of rules accounting for spelling differences resulting from language change over time since the texts of older forms of the Bible in English compared to their modern English equivalents can be quite different (e.g. the old form “saith” versus the modern day form “say”). These processes can be easily described through the rule *y* → *ith*. A further example from the *British Library Microsoft archive*, compensates for English texts from the 15th century, where the suffix “-ynge” has since been replaced by the modern suffix “-ing” in words such as “accordynge” → “according”. Table 5, summarises some of the rules constructed for the *Bible*. The related queries are generated by replacing the characters of the original query with the output of an associated rule, and submitted to the suffix tree.

In summary, the *Type I related queries* capture the processes of deletion, substitution, and insertion, which account for the majority of string permutations that are likely to be found in natural language data, and the *Type II related queries* allow researchers to identify related queries which are not easily identifiable through the *generalised edit distance* approach adopted by the *Type I related queries*.

Table 5. *Type II related queries*: a small set of character rules for the *Bible* represented by old spelling variants extracted from the *Tyndale* and *Wycliffe* Bibles.

Modern English	Early Modern English	Examples
y	ith	say → <i>saith</i>
v	th, st	have → <i>hathe</i> .
⟨word-final⟩	th, yst	mean → <i>meanyst</i> .

5 CASE STUDIES

This section presents several case studies demonstrating the flexibility of the search tools to a range of digital archives that differ in terms of language, domain, quality, and media types.

5.1 UK Medical Heritage Library - Wellcome Trust

“Paratyphoid” is a type of *Salmonella* bacterial infection, which exhibits highly similar symptoms to “typhoid” fever. A researcher searching for documents in Samtla related to Paratyphoid will also retrieve the set of documents for the related infection “typhoid” as it appears as a suffix of the original query. The word-level representation for the documents would require the researcher to submit two separate queries, whereas the character-based approach retrieves both variants simultaneously. This flexibility in querying is also apparent when an archive contains languages from the same language group (e.g. languages of Indo-European including English, French, Spanish, and Italian). These languages may share a common morpheme such as those derived from Greek and Latin, which are commonly found in vocabulary related to disease in the medical domain.

Furthermore, the majority of authors appear to prefer the hyphenated form “Para-typhoid”. Scrolling down the search results reveals that authors also use the unhyphenated form. Author preferences can complicate word-level models, which require preprocessing involving heuristic assumptions on the best approach for document text normalisation.

Report of the Medical Officer of Health for Chislehurst
 ...n 65 Tuberculosis, Mass X-ray and Chest Clinic 65 Typhoid and Para-Typhoid Fever 67 Whooping Cough 68
 STATISTICAL TABLES ...C4128 AC4412(3)(H)SLEH4025 CHISLEHURST AND SDCUP URBAN DISTRICT COUNCIL
 MEDICAL OFFICER OF HE

Report of the Medical Officer of Health for Holborn Borough
 ...THS. The causes of deaths (378) were as follows:- Typhoid and Para-typhoid Fevers 0. Scarlet Fever 0. Whooping Cough
 0. ...IBRARY HOL 37 METROPOLITAN BOROUGH OF HOLBORN. Abridged Report of the Medical Officer of Health for

Report of the Medical Officer of Health for Orpington
 ...n 41 Measles 42 Whooping Cough 42 Tuberculosis 42 Paratyphoid 43 Food Poisoning 44 Dysentery 45 Section H.
 Re ...ursing Service 15 Health Visiting 15 Domiciliary Midwifery 15 Child Welfare Centres 15 Clinics 16 Am

Report of the Medical Officer of Health for Hendon
 ...uring the year:- TABLE II CAUSE MALE FEMALE TOTAL Typhoid and paratyphoid fevers — — — Cerebro-spinal fever 1
 1 2 Scarl ...e war years and on the Doctors, Health Visitors, Midwives and others concerned. The confidence of the public in
 these Services is reflec

Report of the Medical Officer of Health for Hendon
 ...ring the year:- TABLE II. CAUSE MALE FEMALE TOTAL Typhoid and paratyphoid fevers - - - Cerebro-spinal fever 1 2 3
 Scarl

Fig. 4. The search results for the query “paratyphoid” submitted to the *UK Medical Heritage Library*.

5.2 Financial Times newspaper archive

In Figure 5 the researcher is interested in news articles on the bombing of North Vietnam by America during the Vietnam War (1955 - 1975). Scanning the news articles with partial matches for the query, reveals that the researcher could consider reformulating the original query to “*sorties* over North Vietnam”, where the word *sorties* is more prevalent across multiple documents than the term “air strikes”. Furthermore, the fourth document in the search results shows matches for the query for words that are hyphenated, such as “Viet-nam”, which again demonstrates the flexibility of the character-level n -gram representation over the word-level n -gram model, which may not capture this instance without some level of preprocessing.

Easter Road Deaths Up-86 Die in Four Days

..., there was a 10- mile solid jam- accentuated by major road repairs. Bomb shortage American air strikes over Vietnam have been temporarily redu...t to complete the course, and named second overall. were Kenvans Vic Preston and Bob Gerrish (Ford Cortina). Ford won the team prize. Two Honda powered B... temporarily reduced because S. Vietnamese workers at Da Nang have f

Article subtitle: Record Review Copyright holder: The Financial Times Limited. All rights reserved

Tories Will Strive for Debate on Gibraltar

...were sho down during 139 sorties over North Vietnam on Monday, saic U.S. officia... medal total at th Kingston (Jamaica) Common wealth Games rose to 12 as- Charles S...carriers are to concentrate on North Vietnam attacks, said Washington. Br

Copyright holder: The Financial Times Limited. All rights reserved

Vietnam: No U. S. Moves to Escalate the War

...NEWS SUMMARY Vietnam: No U.S. moves to escalate the war American air strikes at missue sites near Hanoi ...y's raids were closer to the N. Vietnam capital than at any time sin...mittee: "I don't think the South Vietnamese will be the first people

Copyright holder: The Financial Times Limited. All rights reserved Illustration: En Ces Temps Noirs De Jactance Et D'incroyance, Notre Dame De La Fin Des Terres Vigilante—from the Mi

Amsterdam Police Fire on Rioters: 30 Hurt

...ation by senior priest Thich Tam Chau. Bombing sorties over North Viet- nam totalled 60 on Monday-an... dropped 12% m. leaflets in South Vietnam, telling North Vietnam troops: Defect or surrende...ion by senior priest Thich TJ- am Chau. Dombing sorties over North Viet- nam totalled 60 on Monday-an

Author titles: Our Industrial Editor Copyright holder: The Financial Times Limited. All rights reserved

No Vietnam Bombing Pause, Says U. S. Unless

...nother pause in its bombing of North Vietnam unless Hanoi (gives an indi...et for U.S. aircraft attacking North Vietnam- besides 105 barges and junks...venth Fleet destroyers shelled North Vietnam shore bat- teries near Dong

Copyright holder: The Financial Times Limited. All rights reserved

Fig. 5. The search results for the query “American air strikes over North Vietnam”, submitted to the *FT newspaper archive*.

5.3 King James Bible

The query “In the beginning god created the heaven and the earth. And the earth was without form, and void; and darkness was upon the face of the deep. And the spirit of god moved upon the face of the waters.”, represents a set phrase from *Luke, Chapter 1*. Researchers of the Bible may submit long set phrases as queries in order to retrieve a known-item. Digital archives such as the *Aramaic Magic Bowl archive* and the *King James Bible* lend themselves to phrase search due to the repetition of religious themes.

The researcher could potentially submit a whole document as the query in order to identify parallel passages that are similar or equivalent across the whole archive of documents.

King James Bible 2003

in the beginning god created the heaven and the earth. and the earth was without form, and void; and darkness was upon the face of the deep. and the spirit of god moved upon the face of the waters. and god said, let there be light: and there ...

Jewish Publications Society Bible 1917

in the beginning g-d created the heaven and the earth. now the earth was unformed and void, and darkness was upon the face of the deep; and the spirit of g-d hovered over the face of the waters. and g-d said: 'let there be light.' an...

King James Bible 2003

...and it came to pass, when men began to multiply on the face of the earth, and daughters were born unto them, that the sons of god saw the daughters of men that they were fair; and they took t

Jewish Publications Society Bible 1917

...and it came to pass, when men began to multiply on the face of the earth, and daughters were born unto them, that the sons of g-d saw the daughters of men that they were fair; and they took t

Jewish Publications Society Bible 1917

...up; for hashem g-d had not caused it to rain upon the earth, and there was not a man to till the ground; but there went up a mist from the earth, and watered the whole face of the ground. then ha

Fig. 6. The search results for the query “In the beginning god created the heaven and the earth. And the earth was without form, and void; and darkness was upon the face of the deep. And the spirit of god moved upon the face of the waters.”, submitted to the *Bible* archive.

5.4 British Library Microsoft archive

The *British Library Microsoft archive* represents a very diverse collection of topics and text genres. Full-text search is not supported in this version of the system, however, the metadata records contain lengthy unstructured text that provides enough useful information about the topics covered by the archive to make the documents discoverable. For example, consider a query “Journal of the Royal Geographical Society”, representing a query submitted by a researcher interested in publications by this journal can locate these documents on the basis of matches in the title, or note field of the metadata. Furthermore, the document ranked third in the search results, retrieved through a partial match for the query, suggests a potentially relevant document published by the *Journal of the Bombay Royal Geographical Society*. Metadata therefore plays an important role when the text of the document content is not reliable due to spelling errors generated by OCR technology, or when the document represents media such as images, and video.

Fig. 7. The search results for the query “the journal of the royal geographical society”; submitted to the *British Library Microsoft archive*.

5.5 Aramaic Magic Bowl archive

The query “Sealed and countersealed are the house and threshold of Dodi bath Ahath from all evil Plagues, from all evil Spirits, and from the Tormentors, and from the Liliths, and from all Injurers, that ye approach not to her, to the house and threshold of Dodi daughter of Ahath.” [44], represents a “parallel passage” repeated across several documents. The name of the client “Dodi daughter of Ahath” appears across several of the documents, which suggests that the client had more than one bowl commissioned on their behalf. The scribes who authored the bowls on behalf of their clients, inserted set phrases from oral history and well-known religious texts such as the Aramaic translation of the Bible. The researchers studying these texts require very tolerant search tools to facilitate the discovery of these set phrases. The search results, in Figure 8, are composed of full and partial matches to the query, which reveal three very similar documents *AIT Bowls CAL 23 AIT 21*, *AIT Bowls CAL 25 AIT 23*, and *AIT Bowls CAL 24 AIT 22*.

AIT Bowls CAL 23 AIT 21

חתיים ומחתם ביתה ואיסקופתה דודי בת אחת מיכל פגעין בישין ומיכל רוחין בישין ומן מבכלתא ומליתא ומיכל זיקין ומזיקין דלא תקרבון לה לביתה ולאסקופתה דודי בת אחת דחתיים ביתלתה עזקין ומחתם בשבעה חתמין ומיכל פגעין...ודי בת אחת דיחיתם ביתלתה עזקין ומחתם בשבעה חתמין ומיכל פגעין בישין ומיכל רוחין בישין ומימבכלתא מן אמן ואמן סלה

AIT Bowls CAL 24 AIT 22

חתיים ומחתם ביתה ואיסקופתה דודי בת אחת מיכל פגעין בישין ומיכל רוחין בישין ומן מבכלתא ומן ליליתא ומיכל זיקין ומזיקין דלא תקרבון לה לביתה ולאסקופתה דודי בת אחת דחתיים ביתלתה עזקין ומחתם בשבעה חתמין#מ.....ם ביתלתה עזקין ומחתם בשבעה חתמין#מין ליליתא ומכל זיקין דלא תקרבון לה לביתה ולאסקופתה דודי בת אחת מן יומה דן ולעלם אמן אמן סלה

AIT Bowls CAL 25 AIT 23

חתיים ומחתם ביתה ואיסקופתה דודי בת אחת מיכל פגעין בישין ומיכל רוחין בישין ומן מבכלתא ומן ליליתא ומיכל זיקין ומזיקין דלא תקרבון לה לביתה ולאסקופתה דודי בת אחת דחתיים ביתלתה עזקין ומחתם בשבעה חתמין#מ.....ם ביתלתה עזקין ומחתם בשבעה חתמין מבכלתא ומיכל זיקין ומזיקין דלא תקרבון לה לביתה ולאסקופתה דודי בת אחת אמן אמן סלה

Hilprecht Bowls Hilprecht 3

חתיים ומחתם ביתה אי ואיסקופתה דאדק בר מחלפתא ולדממי אינתה מן כל פגעין בישין ומן כל רוחין בישין ומן מבכלתא ומן ליליתא ומן כל זיקין ומזיקין ודלא תקרבון לה#לביתה ולאסקופתה דאדק בר מחלפתא דיחיתם ביתלתה עזקין ומחתם ב.....ין ודלא תקרבון לה#לביתה ולאסקופתה דאדק בר מחלפתא דחתיים ביתלתה עזקין ומחתם בשבעה חתמין מן כל פגעין בישין מן כל רוחין בישין מן מבכלתא ומן ליליתא ומן כל זיקין ומזיקין דלא תקרבון לה לביתה ולאסקופתה דאדק בר מחלפתא ולדממי אינתה מן יומה דן וילע

Fig. 8. The search results for a query representing a “parallel passage”, submitted to the *Aramaic Magic Bowl archive*.

5.6 Giorgio Vasari archive

The *Giorgio Vasari archive* supports the research of art history. To illustrate, the researcher in this example is searching for a particular work of art “Madonna with Child”, relating to the Virgin Mary and Jesus Christ, which was a popular theme for artists, who were influenced by the history, culture, and religious teachings in Italy at the time. Figure 9 presents the search results for the English and Italian translations of the documents. Here the researcher is particularly interested in the work of Antonio Veneziano in the year 1380. The metadata search component is useful in this example, as it provides a “bridge” between the two language corpora. Furthermore, documents are also retrieved according to matches in their image captions, enabling the system to locate documents across different media.

<p>Antonio Veneziano</p> <p>... VENEZIANO/ ANTONIO VINIZIANO Madonna and Child . c. 1380, Museum of Fine Arts, Bost...es of the Artists MANY who would fain stay in the country wher</p> <p>Image title: Madonna and Child, c. 1380, Museum of Fine Arts, Bost</p>
<p>Antonio Veneziano</p> <p>De BibliothecaLa biblioteca di BabeleBiblioteca TelematicaCLASSICI DELLA LETTERATURA ITALIANATesti senza diritti d'autore LE VITE DE' PIÙ ECCELLENTI ARCHITETTI, PITTORI, ET SCULTORI ITALIANI, DA CIMABUE INSINO A' TEMPI NOSTRINell'edizione per i tip</p> <p>Image title: Madonna and Child, c. 1380, Museum of Fine Arts, Bost</p>
<p>Margaritone</p> <p>...VASARI'S LIFE OF MARGARITONE Madonna and Child, c. 1270. National Gallery o...ves of the Artists AMONG THE OLD PAINTERS who were much alarme</p> <p>Image title: Madonna and Child, c. 1270. National Gallery of Art, Washington, D.C.HIC JACET ILLE BONUS PICT Current</p> <p>location: I Gallery of Art, Washington, D.C</p>
<p>Margaritone</p> <p>De BibliothecaLa biblioteca di BabeleBiblioteca TelematicaCLASSICI DELLA LETTERATURA ITALIANATesti senza diritti d'autore LE VITE DE' PIÙ ECCELLENTI ARCHITETTI, PITTORI, ET SCULTORI ITALIANI, DA CIMABUE INSINO A' TEMPI NOSTRINell'edizione per i tip</p> <p>Image title: Madonna and Child, c. 1270. National Gallery of Art, Washington, D.C.HIC JACET ILLE BONUS PICT Current</p> <p>location: I Gallery of Art, Washington, D.C</p>
<p>morto da feltre and andrea di cosimo</p> <p>...LTRINI Morto da Feltre. Virgin and Child. Museo Civico, Feltre. LIVES OF MORTO DA FELTRE (1480 circa-1527) and of ANDREA ...ade, which caused him to be held in great estimation, found hi</p> <p>Translation title: Morto da Feltre and Andrea di Cosimo</p>

Fig. 9. The search results for the query “Madonna and Child 1380”, submitted to the *Giorgio Vasari archive*.

6 EVALUATION

The search functionality of Samtla (see Section 4) has been evaluated quantitatively following the standard practice in the IR community, which is the focus of this Section. It is not so easy to perform a quantitative evaluation of the parallel passage discovery performance, as there seems to be no standard performance measure or benchmarking processes for this type of task. Nevertheless, the effectiveness of the discovery functionality has been verified by anecdotal evidence from the different Samtla user groups. To recruit human assessors for the quantitative evaluation of Samtla’s search functionality, we made use of *crowdsourcing* [2, 7, 30, 55] which has proved to be very useful for large-scale labeling or evaluation. The most well-known crowdsourcing platform is probably Amazon Mechanical Turk¹⁷ (*MTurk*), but it is restricted to users resident in the United States only. We finally selected a crowdsourcing platform called Prolific Academic¹⁸. It is part of the Software Incubator at the University of Oxford, and currently has a pool of over 22,000 participants. The document collection used for the evaluation is the King James Bible, because it is relatively easy to find human assessors who are familiar with the content of the Bible. The test set consists of 50 queries of different lengths, from short single-word queries (e.g. “Moses”, and “Jesus Christ”) to longer verbose queries (e.g. “the Lord hath spoken”, and “blessed be the Lord”). We present users with a ranked list of the top-10 search results generated by the Samtla system. Each human assessor has been asked to make relevance judgments on the search results for 40 queries. A search result’s relevance score with respect to the corresponding query would be given at four grades: “not relevant” (0), “somewhat relevant” (1), “quite relevant” (2), or “highly relevant” (3). When the ranked list of search results for a test query were presented to the human assessors to solicit their relevance judgments, they would have been randomly shuffled. This is to avoid the possible overestimation of ranking quality caused by the *presentation bias* aka *position bias* effect (i.e., users tend to assign greater relevance to higher ranked results) [3, 27]. It is challenging to ensure the reliability of relevance judgments obtained from crowdsourcing [17].

Our evaluation task requires the participants to be not only fluent in English but also knowledgeable enough with the corpus (the Bible). Therefore two sanity-check queries (“Jesus Christ”, and “Satan”, had been used to filter out those participants who could not make sensible relevance judgments for such simple cases. Furthermore, some participants may just rush through the evaluation task by randomly assigning relevance scores in order to receive the payment with as little effort as possible, known as “gaming” the system [30]. Therefore the timestamps of each participants’ activities had been used to filter out those who completed the evaluation task at an unreasonably fast speed (e.g. two seconds per search result). This evaluation task attracted 65 participants, the majority of which were male between the ages of 20-30 years, and resident or born in North America. Among them, we excluded 31 incompetent participants who either failed the sanity-check queries or were found to be gaming the system, and also excluded 10 participants due to incomplete submissions. In the end, a total of 24 participants successfully met all criteria and provided a fully completed submission, and they constitute the set of human assessors for the evaluation. For each test query, there would be one ranking of search results according to the relevance scores given by each individual human assessor, which is referred to as the *user ranking*. All such user rankings with respect to a test query are aggregated to produce a *consensus ranking*. Specifically, the consensus ranking is generated by summing up the relevance scores from different human assessors for each search result and then sorting the search results in descending order by the total relevance scores. The consensus rankings reflect the “wisdom of the crowd”, which is considered as the “ground-truth” for our evaluation.

The *system ranking*, i.e., the ranking of search results given by the Samtla system (see Section 4), is compared with the consensus ranking in two ways. First, we evaluate the quality of the system ranking directly using IR ranking measures based on the consensus ranking (see Section 6.1). Second, we check whether the system ranking correlates with the consensus ranking similarly as the user ranking does on average: if so, it implies

¹⁷<https://www.mturk.com/>

¹⁸<https://prolific.ac/>

	disc. func. i	disc. func. $\log_2 i$
system ranking	0.981 [0.980 - 0.981]	0.983 [0.983 - 0.984]

Table 6. The average NDCG scores.

that the quality of the system generated ranking is close to a human assessor's (see Section 6.2). We measure the statistical significance of the evaluation results using the non-parametric *bootstrapping* method [13, 49, 51].

6.1 Ranking Measures

The most widely used performance measure for a ranked list with graded-relevance judgments is the *Normalised Discounted Cumulative Gain (NDCG)* [26]. The *NDCG* utilises a discounting function to model user persistence with respect to whether the user will continue to select documents further down the search results: documents appearing later in the ranked list are unlikely to be as useful to the user as those in the top ranks and therefore only a small part of the document relevance score is passed on to the cumulative gain. One simple discounting function reduces the contribution of the relevance score according to the rank position i . Another popular choice for the discounting function is to use a logarithmic reduction factor $\log_2 i$. The Discounted Cumulative Gain (*DCG*) for a ranked list r is formalised as

$$DCG_k(r) = rel_1 + \sum_{i=2}^k \frac{rel_i}{\log_2 i} \quad (9)$$

where k is the length of the ranked list, rel_i is the relevance score at rank position i , and the term $\log_2 i$ is the discounting function (which could be replaced by alternatives such as i). The score *NDCG* is calculated by normalising the *DCG* into the range $[0, 1]$ as follows:

$$NDCG_k = \frac{DCG_k}{IDCG_k} \quad (10)$$

where *IDCG* denotes the ideal *DCG*, i.e., the maximum possible *DCG* obtained by sorting the search results in the descending order of their relevance grade given by the ground-truth (the consensus ranking in our context). Table 6 shows the system ranking's average *NDCG* scores per query (with the 95% confidence intervals estimated by bootstrapping in square brackets), using both discounting functions i and $\log_2 i$. It can be seen that the *NDCG* scores are very close to the perfect score 1. This means that the Samtla system search functionality frequently places the most relevant documents, matching the users query, at higher positions in the ranked list.

6.2 Correlation Measures

The correlation measures for two rankings that we have used are *Spearman's Footrule* [15] and its variant the *M-measure* [4]. They are both non-parametric and related to well-known statistics, including Spearman's ρ and Kendall's τ [16]. Spearman's Footrule, denoted in short as the *Fr-measure*, is calculated by summing over the absolute differences between the rank positions of each document in both ranked lists:

$$Fr(r_1, r_2) = \sum_{i=1}^k |(r_1(i) - r_2(i))| \quad (11)$$

where r_1 and r_2 are assumed to contain the same set of documents, and k is the size of the ranked list ($k = 10$ in our context). The raw *Fr-measure* score can be transformed into a similarity metric $\widehat{Fr}$ in the range $[0, 1]$ by

	$\widetilde{Fr}$	$\widetilde{M}$
system ranking	0.757 [0.756 - 0.759]	0.761 [0.759 - 0.763]
user ranking	0.716 [0.715 - 0.718]	0.737 [0.735 - 0.739]

Table 7. The average correlation scores.

normalising it:

$$\widetilde{Fr}(r_1, r_2) = 1 - \frac{Fr(r_1, r_2)}{Fr_{\max}(k)} \quad (12)$$

where $Fr_{\max}(k)$ is the maximum possible value of Fr for a ranked list of length k . $Fr_{\max}(k) = \frac{1}{2}k^2$ when k is an even number and $Fr_{\max}(k) = \frac{1}{2}(k+1)(k-1)$ when k is an odd number. When users receive the ranked list of search results, they pay more attention to the documents at higher positions. The M -measure [4] takes this into account when comparing two rankings by giving higher weights to the differences at higher positions as follows:

$$M(r_1, r_2) = \sum_{i=1}^k \left| \frac{1}{r_1(i)} - \frac{1}{r_2(i)} \right|, \quad (13)$$

where r_1 and r_2 are assumed to contain the same set of documents, and k is the size of the ranked list ($k = 10$ in our context). The raw M -measure score can be transformed into a similarity metric $\widetilde{M}$ in the range $[0, 1]$ by normalising it through:

$$\widetilde{M}(r_1, r_2) = 1 - \frac{M(r_1, r_2)}{M_{\max}(k)} \quad (14)$$

where $M_{\max}(k) = \sum_{i=1}^k \left| \frac{1}{i} - \frac{1}{k-i+1} \right|$ is the maximum possible value of M for a ranked list of length k .

The resulting similarity metrics $\widetilde{Fr}$ and $\widetilde{M}$ take on the value 0 for completely uncorrelated rankings, and a value 1 for fully correlated rankings, i.e., identical rankings. Table 7 shows the average correlation scores between the system ranking and the consensus ranking per query (with the 95% confidence intervals estimated by bootstrapping in square brackets), using both $\widetilde{Fr}$ and $\widetilde{M}$. It can be seen that the system ranking correlates well with the consensus ranking, which confirms the effectiveness of the system ranking. Furthermore, Table 7 includes the average correlation scores between the user ranking and the consensus ranking per query, which are also very high implying that the human assessors had a high degree of inter-rater agreement (aka concordance) with each other in terms of their relevance judgments. More importantly, comparing the system ranking and the user ranking (given by individual human assessors), we can see that their correlation scores with regard to the consensus ranking ground-truth are very close (with the system ranking being slightly better). This suggests that the effectiveness of the Samtla system's search functionality is on par with the human-level performance.

7 CONCLUSIONS

The main research contribution of this paper is a novel combination of character-based n -gram Statistical Language Models, space-optimised suffix tree, generalised edit distance, and local sequence alignment with applications to search and comparison of documents stored in digital archives. The techniques complement each other and work well together to provide a domain and language-independent digital infrastructure for the search and discovery of parallel passages of importance to historians and linguists researching cultural contexts recorded in the documents. The use of character-based n -grams stored in a suffix tree data structure enables any archive of documents to be indexed with little preprocessing, making it particularly suitable for large-scale archives containing documents in a single language, multiple languages, and documents spanning several periods and subject to language change over time. Furthermore, reducing the documents and the query to variable length

character n -grams preserves the dependency between the constituents of the “parallel passages”, unlike many word-based retrieval systems constructed from a bag-of-words model. Another advantage is that the matching algorithm can extract both complete words, and word-internal features providing a very flexible query language that is tolerant to variations in vocabulary choice, and spelling errors in the documents and query. In addition, the suffix tree lends itself to non-traditional information retrieval tasks, such as related query generation based on the properties of natural language recorded in the documents, and text mining (discussed in a companion paper to follow). Statistical Language Models provide the mechanism for scoring and subsequently ranking the documents in a well-motivated and relatively simple way. The digital infrastructure makes the Samtla system “future-proof” in the sense that it takes almost no effort to make the system work on a new archive from a different domain or language, complete with the same set of tools and functionality. This enables Samtla to be deployed quickly, and thus helps humanities researchers to cope with the large volume of text data produced by digitisation projects and also unlock more culture heritage archives to the general public. Although, Samtla is still undergoing development, it has already been successfully deployed to support a range of corpora in a number of languages including Aramaic, Syriac, Mandaic, Hebrew, English, German, French, Hungarian, Italian, and Russian. The full Samtla system provides a number of other functionalities (such as document comparison, and leveraging named entities), which will be presented in a companion paper to follow.

For future work, we will try to go beyond lexical matching and leverage the semantics of documents in Samtla by extending word embedding [42] to variable-length *phrase embedding* and extending Key Word In Context (KWIC) [39] to *phrase-in-context* models. Furthermore, we will consider extending the support of Boolean logic operators (such as negation) in query processing [38] from document metadata only to full-text document content as well.

REFERENCES

- [1] 2016. Discussion on the meaning of the Aramaic related queries identified by *Samtla*. Personal communication. (July 2016).
- [2] Omar Alonso and Ricardo A. Baeza-Yates. 2011. Design and Implementation of Relevance Assessments Using Crowdsourcing. In *Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011, Dublin, Ireland, April 18-21, 2011. Proceedings*. 153–164. https://doi.org/10.1007/978-3-642-20161-5_16
- [3] Judit Bar-Ilan, Kevin Keenoy, Mark Levene, and Eti Yaari. 2009. Presentation Bias is Significant in Determining User Preference for Search Results — A User Study. *Journal of the American Society for Information Science and Technology (JASIST)* 60, 1 (2009), 135–149. <https://doi.org/10.1002/asi.20941>
- [4] Judit Bar-Ilan, Mazlita Mat-Hassan, and Mark Levene. 2006. Methods for Comparing Rankings of Search Engine Results. *Computer Networks* 50, 10 (2006), 1448–1463.
- [5] Alistair Baron, Paul Rayson, and Dawn Archer. 2009. Automatic Standardization of Spelling for Historical Text Mining. In *Proceedings of Digital Humanities 2009*. University of Maryland, MD, USA.
- [6] Alistair Baron, Paul Rayson, and Dawn Archer. 2009. The Extent of Spelling Variation in Early Modern English. In *Proceedings of the 30th International Computer Archive of Modern and Medieval English Conference (ICAME)*. Lancaster University, UK.
- [7] D. C. Brabham. 2008. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence* 14, 1 (2008), 75.
- [8] S.F. Chen and J. Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech and Language* 13, 4 (1999), 359–394.
- [9] David Chiang. 2005. A Hierarchical Phrase-based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL '05)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 263–270. <https://doi.org/10.3115/1219840.1219873>
- [10] Y. Choueka. 1989. Responsa: A Full-text Retrieval System with Linguistic Processing for a 65-million Word Corpus of Jewish Heritage in Hebrew. *Data Eng.* 12, 4 (November 1989), 22–31. <http://dl.acm.org/citation.cfm?id=86213.86220>
- [11] Ryan Cordell. 2013. “Taken Possession of”: The Reprinting and Reauthorship of Hawthorne’s “Celestial Railroad” in the Antebellum Religious Press. *Digital Humanities Quarterly (DHQ)* 7, 1 (2013).
- [12] W. Bruce Croft, Robert Cook, and Dean Wilder. 1995. Providing Government Information on the Internet: Experiences with THOMAS. In *In The Second International Conference on the Theory and Practice of Digital Libraries*.
- [13] A. C. Davison and D. V. Hinkley. 1997. *Bootstrap Methods and their Application*. Cambridge University Press. <http://dx.doi.org/10.1017/CBO9780511802843>

- [14] M.J. de Jong. 2007. *Isaiah Among the Ancient Near Eastern Prophets: A Comparative Study of the Earliest Stages of the Isaiah Tradition and the Neo-Assyrian Prophecies*. Brill. <https://books.google.co.uk/books?id=TPAQQAAlAAJ>
- [15] P. Diaconis and R. L. Graham. 1977. Spearman's footrule as a measure of disarray. *Royal Statistical Society Series B* 32, 24 (1977), 262–268.
- [16] Ronald Fagin, Ravi Kumar, and D. Sivakumar. 2003. Comparing top k lists. In *Proceedings of the ACM-SIAM Symposium on Discrete Algorithms*. 28–36.
- [17] H. Garcia-Molina, M. Joglekar, A. Marcus, A.G. Parameswaran, and V. Verroios. 2016. Challenges in Data Crowdsourcing. *Knowledge and Data Engineering, IEEE Transactions on PP*, 99 (2016), 1–1. <https://doi.org/10.1109/TKDE.2016.2518669>
- [18] F. Gibbs and T. Owens. 2012. Building Better Digital Humanities Tools: Toward Broader Audiences and User-Centered Designs. *Digital Humanities Quarterly* 6, 2 (2012).
- [19] Jeff Gill. 2014. *Bayesian methods: A social and behavioral sciences approach*. Vol. 20. CRC press.
- [20] D. Gusfield. 1997. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press.
- [21] Martyn Harris, Mark Levene, Dell Zhang, and Dan Levene. 2014. The Anatomy of a Search and Mining System for Digital Humanities. In *Proceedings of the 2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL)*. London, UK, 165–168.
- [22] Martyn Harris, Mark Levene, Dell Zhang, and Dan Levene. 2016. The Anatomy of a Search and Mining System for Digital Archives. *CoRR abs/1603.07150* (2016). <http://arxiv.org/abs/1603.07150>
- [23] Marti A. Hearst. 2008. Uls for Faceted Navigation: Recent Advances and Remaining Open Problems. In *in the Workshop on Computer Interaction and Information Retrieval, HCIR 2008*.
- [24] Leah Hoffmann. 2013. Looking Back at Big Data. *Communications of the ACM (CACM)* 56, 4 (2013), 21–23.
- [25] Samuel Huston and W. Bruce Croft. 2010. Evaluating Verbose Query Processing Techniques. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '10)*. ACM, New York, NY, USA, 291–298. <https://doi.org/10.1145/1835449.1835499>
- [26] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (October 2002), 422–446. <https://doi.org/10.1145/582415.582418>
- [27] Thorsten Joachims, Laura A. Granka, Bing Pan, Helene Hembrooke, Filip Radlinski, and Geri Gay. 2007. Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search. *ACM Transactions on Information Systems (TOIS)* 25, 2 (2007).
- [28] H. Juusola and Suomen Itämainen Seura. 1999. *Linguistic Peculiarities in the Aramaic Magic Bowl Texts*. Finnish Oriental Society. <https://books.google.co.uk/books?id=I0FZAAAAMAAJ>
- [29] Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-Aware Neural Language Models. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI)*. Phoenix, AZ, USA, 2741–2749.
- [30] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing User Studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '08)*. ACM, New York, NY, USA, 453–456. <https://doi.org/10.1145/1357054.1357127>
- [31] Lauren F. Klein, Jacob Eisenstein, and Iris Sun. 2015. Exploratory Thematic Analysis for Digitized Archival Collections. *Digital Scholarship in the Humanities (DSH)* 30, Suppl-1 (2015), i130–i141.
- [32] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions (ACL '07)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 177–180.
- [33] A. Leff and J.T. Rayfield. 2001. Web-application development using the Model/View/Controller design pattern. In *Enterprise Distributed Object Computing Conference, 2001. EDOC '01. Proceedings. Fifth IEEE International*. 118–127.
- [34] M. Levene. 2010. *An Introduction to Search Engines and Web Navigation* (2nd ed.). John Wiley & Sons, Hoboken, New Jersey.
- [35] VI Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady* 10 (1966), 707.
- [36] Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. Portorož, Slovenia.
- [37] Xiaoyong Liu and W. Bruce Croft. 2005. Statistical language modeling for information retrieval. *Annual Review of Information Science and Technology* 39, 1 (2005), 1–31. <https://doi.org/10.1002/aris.1440390108>
- [38] C. Manning, P. Raghavan, and H. Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.
- [39] Chris Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA.
- [40] Paul McNamee and James Mayfield. 2004. Character N-Gram Tokenization for European Language Text Retrieval. *Inf. Retr.* 7, 1-2 (January 2004), 73–97. <https://doi.org/10.1023/B:INRT.0000009441.78971.be>
- [41] Sean Meyn and Richard L. Tweedie. 2009. *Markov Chains and Stochastic Stability* (2nd ed.). Cambridge University Press, New York, NY, USA.
- [42] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems 26: Annual Conference on Neural Information Processing Systems (NIPS)*. Lake Tahoe, NV, USA, 3111–3119.

- [43] Stefano Mizzaro. 1997. Relevance: The Whole History. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE* 48 (1997), 810–832.
- [44] J.A. Montgomery. 1913. *Aramaic Incantation Texts from Nippur*. Univ. Museum. <https://books.google.co.uk/books?id=qg0TAAAYAAJ>
- [45] Kevin P. Murphy. 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press.
- [46] Geoffrey Rockwell and Stéfán Sinclair. 2016. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. MIT Press.
- [47] T. Rommel. 2007. Literary Studies. In *A Companion to Digital Humanities*. Blackwell Publishing Ltd, 88–96.
- [48] Ronald Rosenfeld. 2000. Two decades of statistical language modeling: Where do we go from here?. In *Proceedings of the IEEE*, Vol. 88. 1270–1278.
- [49] Tetsuya Sakai. 2006. Evaluating Evaluation Metrics Based on the Bootstrap. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '06)*. ACM, New York, NY, USA, 525–532. <https://doi.org/10.1145/1148170.1148261>
- [50] Marcel H. Schulz, Sebastian Bauer, and Peter N. Robinson. 2008. The generalised k-Truncated Suffix Tree for time-and space-efficient searches in multiple DNA or protein sequences. *IJBRA* 4, 1 (2008), 81–95. <http://dblp.uni-trier.de/db/journals/ijbra/ijbra4.html#SchulzBR08>
- [51] Mark D. Smucker, James Allan, and Ben Carterette. 2007. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management (CIKM '07)*. ACM, New York, NY, USA, 623–632. <https://doi.org/10.1145/1321440.1321528>
- [52] D.F. Strauss and G. Eliot. 1860. *The Life of Jesus: Critically Examined*. Number v. 1 in The Life of Jesus. C. Blanchard. <https://books.google.co.uk/books?id=RmdLqnfw1OgC>
- [53] M. Sweetnam, M. Agosti, N. Orio, C. Ponchia, C. Steiner, E. Hillemann, M. Ó Siochrú, and S. Lawless. 2012. User Needs for Enhanced Engagement with Cultural Heritage Collections. In *Proceedings of the International Conference on Theory and Practice of Digital Libraries (TPDL)*. 64–75.
- [54] Kenji Yamada and Kevin Knight. 2001. A Syntax-based Statistical Translation Model. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics (ACL '01)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 523–530. <https://doi.org/10.3115/1073012.1073079>
- [55] Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing Translation: Professional Quality from Non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1 (HLT '11)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 1220–1229. <http://dl.acm.org/citation.cfm?id=2002472.2002626>
- [56] ChengXiang Zhai. 2008. Statistical Language Models for Information Retrieval: A Critical Review. *Found. Trends Inf. Retr.* 2, 3 (March 2008), 137–213. <https://doi.org/10.1561/1500000008>
- [57] C. Zhai. 2009. *Statistical Language Models for Information Retrieval*. Morgan & Claypool Publishers, San Francisco.
- [58] Chengxiang Zhai and John Lafferty. 2001. The Dual Role of Smoothing in the Language Modeling Approach. In *Proceedings of the Workshop on Language Models for Information Retrieval (LMIR) 2001*. 31–36.
- [59] Chengxiang Zhai and John Lafferty. 2004. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Trans. Inf. Syst.* 22, 2 (April 2004), 179–214. <https://doi.org/10.1145/984321.984322>
- [60] Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level Convolutional Networks for Text Classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems (NIPS)*. Montreal, Quebec, Canada, 649–657.