Query Representation for Cross-Temporal Information Retrieval

Miles Efron
Graduate School of Library and Information Science
University of Illinois, Urbana-Champaign, USA
mefron@illinois.edu

ABSTRACT

This paper addresses the problem of long-term language change in information retrieval (IR) systems. IR research has often ignored lexical drift. But in the emerging domain of massive digitized book collections, the risk of vocabulary mismatch due to language change is high. Collections such as Google Books and the Hathi Trust contain text written in the vernaculars of many centuries. With respect to IR, changes in vocabulary and orthography make 14th-Century English qualitatively different from 21st-Century English. This challenges retrieval models that rely on keyword matching. With this challenge in mind, we ask: given a query written in contemporary English, how can we retrieve relevant documents that were written in early English? We argue that search in historically diverse corpora is similar to cross-language retrieval (CLIR). By considering "modern" English and "archaic" English as distinct languages, CLIR techniques can improve what we call *cross-temporal IR* (CTIR). We focus on ways to combine evidence to improve CTIR effectiveness, proposing and testing several ways to handle language change during book search. We find that a principled combination of three sources of evidence during relevance feedback yields strong CTIR performance.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.7 [**Digital Libraries**]: Systems Issues

Keywords

Information retrieval, temporality, digital libraries, book search

1. INTRODUCTION

Today, digitized book collections are growing in size and scope. This raises new challenges for information retrieval (IR) research. Repositories such as Google Books[1] and the Hathi Trust[2] collect millions of scanned volumes with the aim of improving readers' access to the range of information stored in academic libraries. To meet this goal, these collections must support appropriate search interactions.

This paper tackles a specific problem in book search: long-term language change. Though many books in, say, the Hathi Trust are written in English, 20th-Century authors used English very differently than authors did during the 14th- or even the 17th-Centuries. This presents a challenge to retrieval models that rank documents based on the distribution of query words observed in documents. In Google Books, a query term *work* may imply a user's interest in *werk, woork, wyrke* or even *swyinkinge*. All textual IR suffers from the vocabulary mismatch problem, where searchers and authors use different terms for similar ideas. But in historically diverse collections, this mismatch is exacerbated by centuries of linguistic evolution.

Our aim in this paper is to allow a query written in contemporary English to retrieve documents written in the vernacular of older centuries. Thus a person interested in the pedigree of the proverb *many hands make light work* will be able to find a broad range of variants on this theme, from a range of historical periods using a single query. This task is similar to cross-language information retrieval (CLIR), and so we will refer to it as cross-temporal retrieval (CTIR).

This paper's main contribution is a novel approach to CTIR. We present a model that imagines contemporary queries as translations of archaic statements. Given a query in modern English Q^M we attempt to recover terms from the "true" archaic version Q^A. The model yields a distribution over terms that fits naturally into a structured query for the inference network retrieval model. While we describe and test several techniques, the paper's main contribution (described in Section 6.4.2) is a relevance feedback method that combines information from a bilingual dictionary, orthographic evidence and feedback probabilities in a principled way.

2. EXAMPLE USE CASE

Though book search has a strong research community – especially in the INEX book track – many issues that impact the effectiveness of IR in digitized book collections have yet to be articulated clearly. In particular, realistic use cases and their attendant demands on a system are non-obvious for search over digitized book collections. Who would use these collections, and for what purposes? What would realistic queries look like? What constitutes relevance in peoples' interactions with a scanned and OCR'd library?

[1] `http://books.google.com`

[2] `http://hathitrust.org`

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.
Copyright 2013 ACM 978-1-4503-2034-4/13/07 ...$15.00.

In this paper, we limit our attention to a fairly narrow hypothetical use case. Our imagined user – probably a humanities scholar, perhaps a general-interest reader – is interested in a particular phrase or historically durable idea. Examples of such interests include:

- A phrase from a literary work
- A proverb or saying
- A famous literary device such as a metaphor.

Thus our user might be interested in finding variants of the opening lines of Chaucer's *Canterbury Tales*. These lines vary from manuscript to manuscript and edition to edition, each with different spellings and vocabulary. They were also adapted into increasingly modern English as critics and writers influenced by Chaucer echoed his text. To get a sense of the range of these lines' deployment in literature over the years, a searcher would need to use a typical IR system very carefully, relying on a great deal of expert knowledge.

In other cases, there is no canonical text to retrieve. The proverb *many hands make light work* has changed over the years; it has no "true" version. *A communitie maketh the werk lesse* is clearly related to this proverb, though neither rendition is in any sense more authoritative than the other.

We imagine that a user of our system is a person with an interest in finding historical variants of a textual seed which we call the searcher's *exemplar*. We aim to retrieve items that contain *renditions* of that exemplar – appropriations, variants, in-line quotations, etc. This suggests that topical relevance is not a suitable criterion for optimizing our system. Searches like those described here seek variants of an exemplar, not necessarily documents that are about that exemplar. A critic's quote of Chaucer's opening lines would be relevant to this searcher. But so would a 15th-Century play by an author who borrows from *The Canterbury Tales* in the course of his own work. We will operationalize this notion of relevance when we describe our experimental evaluation.

Though this use case is narrow, we believe that it is realistic. It speaks to the diachronic focus that has driven much earlier work on information access in scanned book repositories, [3, 27, 28].

3. PREVIOUS WORK

To the best of our knowledge, the cross-temporal IR problem has not been studied before. However, cognate problems have seen a good deal of attention. Much of this earlier work relates to this paper, both in terms of the problems we approach and the ways we propose solving them.

Historical linguists have lent sophistication and rigor to the study of language change. Most notable in our context is Edward Sapir's theories on "language drift," the process by which a language changes over time to such an extent that earlier texts become unintelligible to contemporary readers [30]. However, although historical linguistics could surely help with the problem of CTIR, its focus on the analysis of linguistic genealogies is only indirectly useful here.

In domains closer to IR, work on document date identification relates to our study. In [11], [17], and [5], variations on language modeling have been brought to bear on the problem of identifying when a given document was written. As in these studies, our approaches to CTIR are based on language models, resting on the assumption that resolving differences among distant centuries' vernaculars is tractable via a high-level estimation of word probabilities.

From a conceptual and practical standpoint, CTIR is similar to two active IR research areas: cross-language IR and IR over noisy text (typically due to OCR errors).

If we consider archaic English and Modern English to be two distinct languages, then the CTIR problem as we have presented it *is* a cross-lingual retrieval problem. Certainly, CLIR methods as surveyed in [10] have a role to play in CTIR. For instance, reliance on dictionary-based and corpus-based translation tools enter into our analysis. Additionally, the approach to structuring queries that we adopt is common in the CLIR literature (cf. [4]). Efforts to solve the vocabulary mismatch problem have brought CLIR methods into monolingual retrieval, as well. Work such as [2] and [12] considers query-document matching as an English-to-English translation process.

The CTIR problem also echoes challenges faced by retrieval over modern OCR'd text. Of course, much of the research on IR for scanned documents is situated in the domain of cultural heritage [6], as is our work. But the similarity is more substantive that this. Foundational work such as [8] presents n-gram methods for supporting search over degraded texts. Additionally, variants of the n-gram methods that followed this work (particularly in the TREC confusion track) will form one of our experimental baselines in Section 8. Of particular interest is [18], where Lam-Adesina et al. note that many IR methods are quite robust against OCR-introduced noise; but relevance feedback is more brittle. Our feedback technique presented in Section 6.4.2 was developed in part due to this observation.

Finally, the task of CTIR is not constrained to search over digitized books. But these collections do make the problem of language drift particularly keen. A good deal of work in the INEX book track has treated OCR-related problems in book search (cf. [13] for an overview). However, the focus of INEX topics have left linguistic change as a relatively minor problem in most previous studies of book search.

Perhaps the most important fact about the literature on book search is that while digitized book data is plentiful, what people will want to do with these data is still largely unknown. We intend this paper to complement recent efforts to expand the scope of interaction design for large book repositories [14, 15, 16].

4. CROSS-TEMPORAL INFORMATION RETRIEVAL FOR BOOK SEARCH

Whatever the nature of users' information needs, repositories of digitized books present a novel challenge for IR systems: massive language drift. A corpus such as Google Books or the Hathi Trust contains texts that are nominally written in the same language (e.g. English), but whose vocabularies are nonetheless dissimilar due to long-term language change. The first line of the famous poem *Sir Gawain and the Green Knight* [32] reads: *Sithen the sege and the assaut watz sesed at Troye*, which translates into contemporary English as *Since the siege and the assault was ceased at Troy.* Seeing this equivalence is easy for scholars of Middle English, but it is difficult for many English speakers, let alone for a computational system.

In this paper, our goal is to support a query issued in contemporary English, retrieving documents that are relevant but that are written in the vernacular of earlier centuries. For convenience, we refer to the user's language as *modern* English, with all old forms of English called *archaic*. This is a simplification in the sense that "modern" has technical implications in certain disciplines of the humanities. Likewise, our umbrella term "archaic" refers to many eras such as Middle English and early modern English. While operationally, we have divided English into two classes – *modern* and *archaic* – the fluidity between these classes will be apparent later in our discussion.

5. QUERY STRUCTURE AND RETRIEVAL MODEL

Though not strictly necessary, in this paper we rely on the inference network retrieval model implemented in the Indri search engine [24, 31] . Given a query Q and a document D, the inference network allows us to compute $P(\mathcal{I}|Q, D, \Theta)$ where $\mathcal{I}$ is the information need that generated Q and Θ contains the parameters of the underlying joint distribution over indexing features. The inference network approach is helpful here because it admits operations on structured queries. The belief operators in the Indri query language have been widely used for cross-language IR in a way that we adopt directly. Consider the user-supplied modern query Q^M, *april showers*. For cross-temporal IR, we might rewrite this as Q^S, `#combine(#wsyn(0.5 april 0.3 aprile 0.2 aprylle) #wsyn(0.7 showers 0.2 shours 0.1 rain))` where the `wsyn` operator treats word variants as weighted synonyms. Of course, finding proper synonyms and weights is hard; it is the task that comprises the contribution of this paper.

Given a modern query $Q^M = \{m_1, m_2, \ldots, m_n\}$, we will build a structured query Q^S that consists of n `wsyn` clauses such that the i^{th} clause contains k archaic translations of the i^{th} modern term. For retrieval we use the "Boolean and" (`band`) belief operator to calculate

$$P(\mathcal{I}|Q, D, \Theta) = \prod_{i=1}^{n} \sum_{j=1}^{k} P(a_j|D)^{w_{ij}} \tag{1}$$

where we have k translations for each of our n observed terms and w_{ij} is the weight of the j^{th} translation for the i^{th} observed term. We use Indri's default parameters such that the probabilities are estimated using Dirichlet smoothing of multinomial language models with a hyperparameter of $\mu = 2500$. In other words $\Theta = \{\mu P(t_1|C), \mu P(t_2|C), ..., \mu P(t_V|C)\}$ where $P(t|C)$ is the probability of term t in the collection and V is the size of the vocabulary.

6. QUERY TRANSLATION

Given the retrieval model given by Eq. 1, a core challenge of CTIR lies in finding archaic synonyms for each modern query term and assigning weights to them. To accomplish this, we pose the problem as a variation on cross-language retrieval, where modern and archaic English are considered to be two distinct languages. We assume that a modern query Q^M is generated by an underlying information need $\mathcal{I}$. Our goal is to rephrase Q^M as a query Q^S that will retrieve documents useful to $\mathcal{I}$ but expressed in a way that will retrieve documents written in archaic English[3].

We begin by assuming that an observed modern query Q^M is a translation of an unknown archaic query Q^A. Our goal is to recover Q^A from Q^M. To simplify this process we further assume that an observed query word m_i translates to 0 or more archaic terms $\{a_{i1}, a_{i2}, \ldots, a_{ik}\}$. In this paper we estimate the probability of each archaic translation of a modern term m_i in isolation from all other observed terms $m_j \neq m_i$, leaving non-independencies for future work.

For a modern term m_i, we can rank the V terms in the vocabulary $\{a_1, a_2, \ldots, a_V\}$ in decreasing order of $P(a|m_i)$. Those archaic terms whose $P(a|m_i)$ exceeds some threshold τ are added as a synonym for m_i in the final query, with $P(a|m_i)$ as their weight in the `wsyn` clause.

For the modern query $Q^M = \{m_1, m_2, \ldots, m_n\}$ we choose elements for the `wsyn` query clause associated with term m_i by taking the top k terms from $P(a|m_i)$. This gives the clause: `#wsyn(`$P(a_{i1}|m_i)\, a_{i1}\ P(a_{i2}|m_i)\, a_{i2}\ \ldots\ P(a_{ik}|m_i)\, a_{ik}$`)`.

Our final query has an analogous clause for each of the n observed modern query terms. For simplicity we assume that all observed terms (i.e. all `wsyn` clauses) are given equal weight in the final query.

6.1 Translation in CTIR

All of the models we present are based on a simple probabilistic form, which we present here. Let T be a body of text such as a document, a corpus or a dictionary. We define $P(a|m, T)$, the probability that in a text T, the term a acts as a translation of m. By the definition of conditional probability, we have:

$$P(a|m, T) = \frac{P(a, m, T)}{P(m, T)}. \tag{2}$$

We can think of $P(.|m, T)$ as a multinomial characterized by Θ_m, the probability that m is a translation of $\{a_1, a_2, \ldots, a_V\}$ for a vocabulary of size V. To estimate Θ_m, the maximum likelihood estimator is:

$$\hat{P}^{ml}(a|m, T) = \frac{|a, m, T|}{\sum_{i=1}^{|T|} |t_i, m, T|} \tag{3}$$

where $|a_i, m, T|$ is the number of times that term a_i acts as a translation of m in T.

The following sections will define several approaches to determining if an occurrence of a in T is acting as a translation of m. Thus, Eq. 2 is generic. To make it useful we need a method of calculating the strength of the translational relationship between a and m. For any source of evidence $\mathcal{E}$, we will define an indicator function $\varphi_{\mathcal{E}}(a, m)$ that evaluates to 1 if criteria specific to $\mathcal{E}$ are met by a and m, or 0 otherwise, allowing us to obtain the counts specified in Eq. 3. The precise way to define φ will depend on the type of evidence we are using.

6.2 Dictionary Evidence

Bilingual dictionaries are a mainstay of cross-language IR [1, 20]. Researchers in the humanities have created similar resources that are helpful for cross-temporal IR. Of course many machine-readable dictionaries of Old and Middle English are available[4]. But these are often small, with commentary that makes using them in automated settings difficult.

[3] Of course, a search engine would retrieve both archaic *and* modern documents relevant to $\mathcal{I}$. However, retrieving modern versions is a problem that is at least conceptually distinct from our focus here.

[4] `http://archive.org/details/ElementaryMiddleEnglishGrammar`, `http://www.gutenberg.org/ebooks/10625`, e.g.

However, one resource stands out in this field. The digital humanities MorphAdorner project[5] contains a list of $\langle archaic, modern \rangle$ word pairs. After case-folding this "dictionary" contains 202,285 pairs. The MorphAdorner lexicon is helpful for CTIR due to its high quality. In personal correspondences, project developers explained that compiling the list took substantial work by subject experts. Because it was built by human experts there are very few false positives in the dictionary. Likewise, from its inception, the dictionary was intended to support automated systems. Thus using it requires no disambiguation or other data cleaning.

But like any bilingual dictionary, the MorphAdorner word list presents problems for use during CTIR:

1. **Translation weights.** Word pairs in the list are simple tuples, lacking a measure of translation quality. For instance, the modern term *authority* maps to 46 archaic variants in the list. Some of these terms appear often in archaic English, such as *auctoritee* and *authorite*. But *uthority* is rare. During IR, it would be helpful to let the strength of an *archaic* ↔ *modern* term association inform our query model.
2. **Out of vocabulary terms.** Even a list of 202,285 term pairs will suffer a sparse data problem due to the long-tailed distribution of terms in text.

Despite these issues, the structure and high quality of the MorphAdorner dictionary makes a simple translation indicating function obvious. Given a dictionary D of archaic-modern term pairs:

$$\varphi_D(a, m) = \begin{cases} 1 & \text{if } \langle a, m \rangle \in D \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

Plugging Eq. 4 into Eq. 3 we have the estimator:

$$\hat{P}_D^{ml}(a|m, D) = \frac{|a, m, D|}{|m, D|} \tag{5}$$

where $|m, D|$ is the number of tuples in the dictionary with m as their modern entry, and $|a, m, D|$ is the number of these $|m, D|$ tuples that have a as their archaic entry.

Eq. 5 allows us to specify a "dictionary" query Q^D. For an n-term modern query Q^M, Q^D has $\nu \leq n$ `wsyn` clauses. The elements of each `wsyn` clause are simply the archaic terms with non-zero $\hat{P}_D^{ml}(a|m, D)$, with all archaic terms weighted uniformly, in accordance with Eq. 5.

6.3 Orthographic Evidence

Cross-temporal IR invites us to supplement dictionary-based evidence with a second source of information. Unlike, say, English and French, where cognates are rare, word spellings in 20th-Century English and Early Modern or even Middle English are often similar. Most English readers are familiar with the addition of a terminal *e* when translating a modern term to an archaic version, as in *april* ↔ *aprile*. It is likewise easy to recognize that *aperil* is a "cognate" for *april*. These similarities are common, and thus provide obvious leverage for translating modern terms into archaic terms. We refer to this as *orthographic*, or *spelling* evidence.

Orthographic evidence is attractive because it relieves the problem of out-of-dictionary terms. Spelling evidence might also be useful in determining translation weights for structured queries. However, several problems impinge on using orthographic evidence in cross-temporal IR.

- A few patterns such as adding a terminal e are common in modern-archaic word pairs. But most orthographic distortions are idiosyncratic, making it hard use spelling systematically.
- The risk of false positives is high. Adding a terminal e to a modern word is often safe. But it can easily lead to errors; e.g. adding an e to *man* yields the (incorrect) modern term *mane*.
- Combining dictionary and spelling information is non-trivial. How to combine these sources of evidence in a sensible way is a challenge.

Regardless of these difficulties, we hypothesize that if handled properly, orthographic modern-archaic word similarity should improve the quality of query translations. Our goal in this regard is to induce a model that gives $P_S(a|m)$, the spelling-related probability that archaic word a is a viable translation of modern word m. Intuitively, $P_S(aprile|april)$ should be large, as should $P_S(avril|april)$. But $P_S(peril|april)$ should be small, with $P_S(man|april)$ still smaller.

The appeal of using orthography to identify translations is that it frees us from our dictionary's limitations. Instead of D, we will now use the entire text of the corpus C as T in Eq. 2. That is, we simply consider C to be a very large document. Referring to Eq. 3 we have the spelling probability:

$$\hat{P}_S^{ml}(a|m, C) = \frac{|a, m, C|}{\sum_{i=1}^{V} |t_i, m, C|} \tag{6}$$

where the number of translation events between a and m is determined by a function $\varphi_S(a, m)$ which we define in the following subsection.

6.3.1 Stochastic Edit Distance

To assess whether a and m are "close enough" to be archaic-modern translations with respect to orthography, we need a measure of string similarity. A good deal of prior work has treated this problem. Much of this work focuses on variants of the edit distance. Given given two strings a and m, a set of permissible edit operations, and a cost function ℓ, the edit distance $d(a, m)$ between a and m is the minimum cost of transforming a into m via our allowed operations[6].

If all costs equal 1, then this definition gives the well-known Levenshtein distance. But the Levenshtein distance has two drawbacks. First, it has no ready probabilistic interpretation that would help us integrate it into CTIR. Also, in CTIR, some operations intuitively merit different costs.

Several probabilistically motivated edit distances that address these issues have been proposed in the literature. We rely on results by Oncina and Sebban [26] and Ristad and Yianilos [29]. In particular, we make use of the model developed by Oncina and Sebban which finds the maximum likelihood estimator of the cost matrix via the EM algorithm[7]. Since we use this algorithm without alteration, we omit explicating it here, referring readers to [26, Appendix

[5] `http://morphadorner.northwestern.edu/morphadorner/`

[6] We use the common edit operations, insert, delete and replace.

[7] Our results rely on the open-source SEDIL software, `http://labh-curien.univ-st-etienne.fr/SEDiL/`.

A] for a full treatment. For our purposes, two details of the algorithm are important. First, it operates without any domain knowledge aside from the training tuples. Second, based on a corpus of archaic-modern term tuples, we obtain a translation model $t(a|m)$ by normalizing the results of the EM algorithm such that $t(.|m_i)$ sums to one for all m_i.

We trained $t(a|m)$ on the MorphAdorner dictionary. This leads to the spelling indicator function:

$$\varphi_S(a,m) = \begin{cases} 1 & \text{if } t(a|m) > \tau_S \\ 0 & \text{0, otherwise.} \end{cases} \tag{7}$$

for an empirically determined threshold τ_S which governs how similar a term a must be to m to be considered a viable translation for m.

6.4 Combining Evidence

The previous two subsections introduced sources of evidence that might help cross-temporal IR. But combining these sources would presumably improve effectiveness of CTIR, much as evidence combination has aided CLIR [25]. This section presents two methods of combining dictionary and spelling evidence in the framework given by Eq. 2.

6.4.1 Naive Combination

It is tempting simply to assume that strong evidence on both dimensions – dictionary and spelling – should increase our confidence in a translation. We refer to this approach as *naive combination*.

We can combine dictionary and spelling evidence by using the dictionary model $P_D(a|m,D)$ as the basis for a prior to find the maximum a posteriori (MAP) estimate of the spelling model. Because $P_S(a|m,C)$ is a multinomial, we use the Dirichlet distribution (the multinomial's conjugate prior) with parameters $\{\mu_S P_D(a_1|m,D), \mu_S P_D(a_2|m,D), \dots, \mu_S P_D(a_V|m,D)\}$ as our prior, with a smoothing hyperparameter μ_S.

The maximum *a posteriori* estimate is thus:

$$\hat{P}_N^{map}(a|m,C) = \frac{|a,m,C| + \mu_S P_D(a|m,D)}{\sum_{i=1}^{V} |t_i,m,C| + \mu_S} \tag{8}$$

where $P_N(a|m,C)$ is the "naive" translation model.

6.4.2 Relevance Feedback

Eq. 8 alleviates the constraints of dictionary-based translation, but at the cost of a huge increase in the size of the search space. Because Eq. 8 operates in a completely unsupervised way over the entire vocabulary, it is likely that a query populated with all terms with non-zero probability according to Eq. 8 will contain many false positives.

Thus, we hypothesize that weak orthographic similarity is a strong indicator that a is a bad translation of m. But it is not the case that high orthographic similarity implies a good translation. If this is true, while Eq. 8 allows us to ignore many low-similarity terms, it will still suffer from a high false positive rate.

To overcome this problem, a logical approach is to constrain the search space of translation discovery. Relevance feedback accomplishes this. Given R, a set of k (pseudo-) relevant documents obtained by the dictionary-based query Q^D, we will limit consideration of translation to terms that occur in R, aiming for two effects:

1. **Term identification.** The set of relevant documents may contain translations of m that were not present in the dictionary. We can add these to the final query.
2. **Term weighting.** The dictionary gives deterministic evidence about the relationship between a and m. But with respect to a particular information need, some translations of m may be preferable to others. Feedback lets us update the weights of translations.

The central idea in our feedback approach is that we only allow terms that appear in the pseudo-relevant documents to alter our query model based on orthographic similarity to observed modern query terms.

This approach uses intuition similar to He's work on CLIR [9]. However, our approach is unique in several senses. Aside from the obvious difference in focus (CTIR vs. CLIR) and the attendant differences in evidence, our approach differs from He's mathematically and algorithmically. He used feedback to improve a translation model learned from parallel corpora, linearly interpolating model weights with dictionary probabilities. We have no parallel corpus. This invites a Bayesian approach, considering the dictionary model as a prior over translations which we use to smooth the probabilities estimated during feedback.

In the remainder of this section, we assume that based on our initial, dictionary-built query, we have retrieved $k = 20$ pseudo-relevant documents. We concatenate these k documents into a large pseudo-document R.

Continuing with our earlier notation, we have the relevance feedback probability of a given m:

$$\hat{P}_R^{ml}(a|m,R) = \frac{|a,m,R|}{\sum_{i=1}^{V} |t,m,R|}. \tag{9}$$

Again, we enumerate the translation events between a and m in R via the indicator function:

$$\varphi_R(a,m) = \begin{cases} 1 & \text{if } t(a|m) < \tau_R \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where $t(a|m)$ is the orthographic probability (i.e. string similarity) and τ_R is a threshold that governs how similar a term a in the feedback documents must be to m to be considered a viable translation for m.

As in the case of our naive combination model, in this case we update our feedback probabilities by considering the dictionary model as the basis for a Dirichlet prior: $\{\mu_R P_D(a_1|m,D), \mu_R P_D(a_2|m,D), \dots, \mu_R P_D(a_V|m,D)\}$, giving:

$$\hat{P}_R^{map}(a|m,R) = \frac{|a,m,R| + \mu_R P_D(a|m,D)}{\sum_{i=1}^{V} |t_i,m,R| + \mu_R}. \tag{11}$$

6.5 Summary of Models

We have presented four models for building CTIR queries. Table 1 summarizes them and lists an abbreviated name for each. A few points about these models are worth stressing:

- Orthographic evidence operates as a binary variable via the indicator functions φ_*. That is, a term a is classified as a translation of m if $t(a|m)$ exceeds a threshold τ_*. The thresholds τ_S and τ_R differ because the CFB model searches for translations over an elite set of documents, while SPELL and CNAIVE search all

Table 1: Summary of Baseline and Experimental Query Models. Baseline methods are shown with a gray background, with more novel models shown against a white background.

Name	Definition	Parameters	Description
DICT	$P_D(a, \|m, D)$	none	Simple dictionary lookup of candidate translations. Uniform weights.
USER	NA	unigrams=0.85, #ow1=0.1 #uw8=0.05	Judges cut-and-paste a relevant passage to create a query. Queries formed via sequential dependency model [23].
RM3	NA	k = 20 documents, n = 20 terms, $\lambda = 0.5$ interpolation	Relevance model 3 of [19]. Feedback terms interpolated with Q^D (Eq. 5).
SPELL	$\hat{P}_S^{ml}(a, \|m, C)$	$\log \tau_S$ =-2.1	Includes all terms from the vocabulary whose string probability $t(a\|m) < \tau_S$.
CNAIVE	$\hat{P}_N^{map}(a, \|m, C)$	$\log \tau_S = -4.2$, $\mu_S = 20000.0$	Naive evidence combination. Uses Bayesian updating to improve the spelling model based on dictionary evidence. Considers all terms in the collection. Weights terms by dictionary evidence and collection frequency.
CFB	$\hat{P}_R^{map}(a, \|m, R)$	$\log \tau_R = -9.0$, $\mu_R = 10.0$, $k = 20$ documents	Same as CNAIVE but the set of possible translations is constrained to the vocabulary of k feedback documents. Weights terms by dictionary evidence and frequency in feedback documents.

documents. Thus CFB can tolerate a wider range of spelling variation than SPELL and CNAIVE can.

- The Dirichlet hyperparameters μ_* govern the extent to which we trust dictionary evidence versus spelling evidence observed in texts (either the whole collection or feedback documents).
- CNAIVE and CFB are identical except for the documents they consider – the full corpus versus only feedback documents – and their parameterization.

7. EXPERIMENTAL DATA

To evaluate our approaches, we built a new test collection. We did not use a pre-existing collection such as the INEX book track's corpora due two considerations:

1. We needed to ensure a high degree of historical diversity in our collection.
2. The types of information needs (and queries) needed for our use case were not present in existing collections.

This section details our test collection.

7.1 Documents

Documents came from two sources. Members of the Google Books team (not affiliated with the authors) made the text of 3,690 volumes available to us. These were a random sample of out-of-copyright texts by the authors listed on three Wikipedia pages in October 2010[8]. These samples had many redundancies, with duplicate and near-duplicate pages being common; these were retained. Documents resulted from OCR on scanned pages, leading to typographical noise.

Besides the Google Books data, we harvested 34,806 full-text books from Project Gutenberg in October 2010. These books were selected by requesting all titles labeled as being in English, Middle English, or Old English.

For our experiments, we set the page (as opposed to the book, chapter, etc.) as the unit of retrieval. For the Google Books data, page-breaks were already present. The Gutenberg data, however, were simply long text files. We split these files into non-overlapping passages of no more than 300 words to form retrievable "documents." The 300-word window was intended to mimic the length of book pages, though it was chosen without rigorous estimation.

All metadata was removed from documents, and texts were indexed using Indri, with no stemming or stoplists applied at index time. This yielded an index containing 25,845,101 documents and a vocabulary size of 6,657,631 terms. We did not perform any language identification to learn which documents were indeed modern and which were archaic; we preferred to let our models tackle this problem.

7.2 Topics

Topics were developed by two hired subject experts in digital humanities – one masters and one doctoral student. The students self-identified as users of Google books. As they created queries, topic developers explored the collection using a simple web interface to the index described above.

Each topic developer was given a page-long description of the proposed CTIR use case. Based on this, they were asked to imagine exemplars that they would like to learn about. Developers were free to choose exemplars that are very common (such as famous Biblical passages) or rarer ones. The main imperative was for them to identify exemplars that would plausibly be of interest to digital humanists in the context of CTIR.

For each exemplar, the developer was asked to write a single version of it in modern English. These were used as queries during experimentation. The developers were instructed to choose a modern phrasing that would be a plausible starting point for a searcher interested in the exemplar. Finally, topic authors wrote a description of the exemplar that their query referred to as a basis for relevance judging.

This process raises at least two objectionable points. First, in some sense, our topic development was done backwards

[8] `http://en.wikipedia.org/wiki/List_of_English_writers`, `http://en.wikipedia.org/wiki/Classical_Latin`, `http://en.wikipedia.org/wiki/Classical_Greek`

from a more realistic scenario. Generating modern renditions of a known archaism is artificial. Second, with no "correct" way to write an archaic exemplar as a modern query, subjectivity entered the process of query creation. However, we argue that the first point – artificial order of operations in query-building – is tolerable; developers were asked to consider the overarching problem, and their expertise in the field put them in a good position to handle this ambiguity. As for the latitude in query expression, this is little different than any other topic development, where an abstract information need must be couched in a particular phrasing.

Developers created a total of 53 topics. Two topics were later found to have zero highly relevant documents and were removed. Five topics were chosen with uniform probability to be used for model training. The remaining 46 topics were used for testing.

Queries ranged in length from 3 to 23 words, with a median of 6 and mean 7.65 terms. Figure 1 shows that the MorphAdorner dictionary provides good coverage of the queries, with only six queries having more than one out-of-dictionary term. Based on this, we believe that the simple dictionary-based query is likely to perform well for these data.

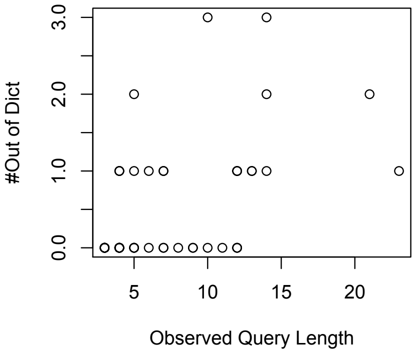

Figure 1: Length of Test Queries and each Query's Number of Out-of-dictionary Terms.

7.3 Relevance Judgments

Unfortunately, the topic authors were not available to serve as relevance assessors. So four Ph.D. students in Medieval Studies were hired to perform relevance judgments. These students were all self-identified experts in early English. Assessors were shown the same topic development guidelines that query authors had read. In addition, assessors were given instructions on how they were to judge document relevance. They then completed an initial trial of 25 judgments, after which they raised any questions before moving on to the bulk of their work.

Judging was done on a three-point scale: 0=not relevant, 1=somewhat relevant, 2=relevant. Assessors were also allowed to say "I don't know," though this option was never used. In Section 8, all effectiveness measures except NDCG treat judgments of 1 and 2 as relevant. NDCG leaves the three-point scale intact.

The criteria for relevance in the context of CTIR are not obvious. Judges were instructed to consult query-document pairs and ask *does the document contain language that intentionally echoes the query*? By *intentionally*, we did not mean that the author needed to know his source. Instead, he needed to know that he was using language that has the history referenced by the query.

Novelty and document quality were not taken into account during the judging process.

Pooling was conducted by running methods DICT, RM3, CNAIVE, and CFB described in Table 1 over the queries (with parameters chosen empirically) . These results were then pooled at a depth of 50 documents per query per model, yielding a total of 8,791 judgments.

Because our assessors were paid domain experts, we only assigned one assessor per judgment, yielding more judgments than multiple assessments per pair would allow. Nevertheless, we wanted some sense of inter-rater reliability. Thus we sampled 50 query-document pairs: 5 queries taken uniformly from our 53, with 10 documents sampled uniformly from those documents that were in at least three of our pools. All assessors rated these query-document pairs. On the three-point relevance scale, this gave a Fleiss kappa agreement of 0.638. Conflating ratings of 1 and 2 to a single *relevant* class, increased kappa to 0.752. Both of these statistics gave $p < 0.001$, suggesting that our assessors largely agreed on the criteria for assigning relevance judgments.

8. EMPIRICAL EVALUATION

Because CTIR is largely an unstudied problem, defining a reasonable baseline for comparing approaches is non-trivial. We define three baselines, shown in the gray rows of Table 1. The run DICT is the simple dictionary lookup described in Eq. 5. The USER run is in some sense an oracle condition; relevance assessors were asked to cut and paste a single, highly relevant passage from a document to form the query verbatim. This text was then represented using the sequential dependency variant of the Markov Random Field model [22]. The USER run is thus not strictly comparable to the others we report because it is based on explicit feedback. But it isn't a pure oracle condition since USER queries stem from a single relevant document and are not expanded in any way. USER queries are high-quality but narrow.

8.1 Omitted Results

Due to space constraints, we do not report results from several models because they were unsuccessful; we list them here for completeness. As additional baselines, we used the character n-gram model described in [8]. We also implemented several ad hoc models based on soundex query transformation (cf. [33]). These methods did not approach the effectiveness of the simple DICT run, so we did not pursue them further. Of course, this is not a defect in the approaches themselves. Rather it simply means that they did not transfer from their original domains to CTIR. Finally, standard Rocchio and BM25 pseudo-feedback were less successful than the reported relevance models.

8.2 Model Effectiveness

Table 2 summarizes the outcomes of our experimental assessment. Results are based on retrievals of 100 documents per query. In the table, statistically significant changes via a permutation test are shown as follows. $\triangle$: improvement over DICT has $p < 0.05$, $\blacktriangle$: improvement over DICT has $p < 0.01$, $\uparrow$: improvement over USER has $p < 0.05$, $\Uparrow$: improvement over USER has $p < 0.01$, $\downarrow$: decline wrt USER has $p < 0.05$, $\Downarrow$: decline wrt USER has $p < 0.01$.

Table 2: Summary of Retrieval Effectiveness. Statistics are mean average precision (MAP), number of relevant retrieved (Rel Ret), R-precision (Rprec) and normalized discounted cumulative gain (NDCG).

Model	MAP	Rel Ret	Rprec	NDCG
DICT	$0.1702^{\downarrow}$	542	$0.1932^{\downarrow}$	$0.2932^{\Downarrow}$
USER	$0.2793^{\triangle}$	554	$0.2944^{\triangle}$	$0.4606^{\blacktriangle}$
RM3	$0.1753^{\Downarrow}$	$532^{\Downarrow}$	$0.195^{\Downarrow}$	$0.2787^{\Downarrow}$
SPELL	$0.1316^{\Downarrow}$	521	$0.1543^{\Downarrow}$	$0.2175^{\Downarrow}$
CNAIVE	$0.1688^{\downarrow}$	$551^{\downarrow}$	0.1961	$0.2869^{\Downarrow}$
CFB	$0.3025^{\blacktriangle\uparrow}$	$870^{\blacktriangle\uparrow}$	$0.3143^{\blacktriangle\uparrow}$	$0.4528^{\blacktriangle}$

A few results are unsurprising: The dictionary-based queries (DICT) do seem like a reasonable baseline. The condition with known relevant text as queries (USER) performs very strongly. And in the bottom three rows, adding increasing amounts of structure to our dictionary-orthography combinations improves effectiveness.

The most obvious point of interest in Table 2 is the strong performance by our orthographically informed feedback method, CFB. Except for the semi-oracle USER run's NDCG, CFB outperforms all methods on all effectiveness measures.

By comparing CFB to DICT and RM3 we can see another interesting result. Pseudo-relevance feedback helps when it is used to alter a query by combining feedback and orthographic evidence via CFB. In contrast, standard feedback did not improve over the simple dictionary method. It is important to note that our implementation of RM3 interpolated the feedback model with the dictionary model; i.e. RM3 it did not give up the query structure of the other models. Also, the RM3 run used a custom stoplist with words added to the standard Indri stoplist to improve performance.

Figure 2 helps us assess the relationship between the two highest-performing runs – USER and CFB. In Figure 2 we see that both methods give higher MAP than the DICT run on most queries. However, CFB is "safer" than the USER run insofar as CFB's queries whose effectiveness decline with respect to DICT do so less than comparable declines seen for USER. This is not surprising, as the USER run puts all of the query burden on the words obtained from a *single* relevant document. Thus USER is very good at retrieving some relevant documents. But it does this at the cost of overfitting the supplied relevant text. Not surprisingly, if we remove the documents from which the judges extracted the query text for USER, its MAP declines to 0.2096.

Though CFB performed well, Table 2 shows that using spelling evidence is risky. The purely orthographic SPELL model was significantly inferior to the simple dictionary lookup. And CNAIVE saw a small, though not significant decline as well. This suggests that there is information in orthographic information, but that using it well is difficult.

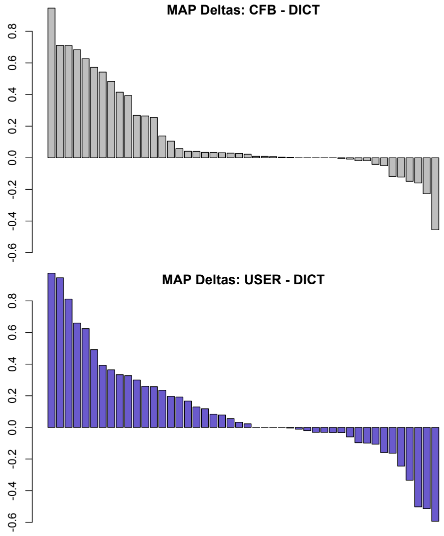

Figure 2: Query-by-Query Difference in MAP between DICT and CFB (Gray) & USER (Blue). Each bar corresponds to one test query. Bar height is MAP under an experimental condition minus MAP of DICT.

8.3 Feedback for CTIR, Further Analysis

The CFB feedback method aims to improve a simple DICT query in two ways: by discovering candidate translations that were not in the dictionary and by learning weights for query terms. The formalism given in Section 6.4.2 dictates how these goals are combined in CFB. But a logical question is: which aspect – term discovery or term weighting – is most responsible for the improvements over DICT seen in Table 2? To pursue this question, we created two variants of CFB that isolate these effects (at the expensive of becoming heuristic in motivation). CFB-W reweights terms from the dictionary query as Eq. 11 dictates but does not add any query terms. Conversely, CFB-T adds all feedback terms to the model that constitute "translations" according to Eq. 10. But in CFB-T, all terms from the dictionary receive a weight of 1.0, while added terms are weighted at 0.5.

Table 3: Retrieval Effectiveness Measures for variants of CFB Feedback Method. Statistically significant changes via a permutation test are shown as follows. ▼: decline wrt CFB has $p < 0.01$.

Model	MAP	Rel Ret	Rprec	NDCG
CFB-W	$0.1603^{\blacktriangledown}$	$523^{\blacktriangledown}$	$0.1817^{\blacktriangledown}$	$0.2932^{\blacktriangledown}$
CFB-T	0.3106	907	0.3176	0.4631

Table 3 suggests that the crucial mechanism for CFB is its ability to discover new terms. If we disable this behavior (yielding CFB-W) performance declines significantly. But if we ignore the induced term weights of CFB (yielding CFB-T) we actually see an increase in effectiveness.

9. DISCUSSION AND FUTURE WORK

Our experiments suggest that methods for handling language change have a strong effect on the success of the types of queries studied in this paper. Without a canonical baseline, it is difficult to say if our methods are performing "well." However, the MorphAdorner dictionary is an unusually large and clean knowledge base by CLIR standards. Thus we argue that the DICT model gives a reasonable baseline.

To improve on this performance, it is logical to enlist orthographic evidence, since it often provides a clear path from a modern term to its historical variants. But our hypothesis that orthographic evidence can only help CTIR in a limited way was borne out experimentally. If a term a is orthographically similar to m, that does not imply that a is a good substitute for m in a CTIR query. But it is also the case that if we fail to see orthographic similarity between a and m, it is unlikely that we have a match that we should trust without further corroboration.

We anticipated a benefit from using feedback to moderate string similarity's influence during retrieval. The strength of the CFB model bears out this hypothesis. By constraining the domain of possible orthographic matches to those in high-quality documents, we avoid the high false positive problem that the purely string-based approach such as SPELLING encountered.

Several limitations of this study are worth noting. We omitted discussion of the sensitivity of the parameters μ_* and τ_* because sweeps on our training queries showed little interesting change. But the poor performance of the CNAIVE model suggests that closer attention to parameterization would be helpful. More importantly, using only five training queries limited our ability to assess parameterization robustly.

Another limitation hinges on temporal diversity. Between "archaic" and "modern" English there is a spectrum of vernaculars. Effective CTIR should retrieve documents across this spectrum. But without a clear metric for assessing such diversity, we have not reported our success in finding documents from diverse periods.

Finally, a fourth type of evidence could be used for CTIR: translation models learned from parallel corpora. We built such a model using Biblical translations, but found that the high number of out-of-vocabulary query terms made its use infeasible. Perhaps an approach based on the less restrictive notion of comparable corpora will allow us to incorporate such evidence. In future work we plan to pursue this.

Two other avenues will inform our future work on CTIR. First, Table 3 suggests that our feedback model performs well but not optimally. Though probabilistically convenient, Eq. 11 is not an optimal way to combine dictionary, spelling and feedback information. In future work we will pursue how to exploit the strengths of CFB more fully.

Our second avenue for future work lies in expanding the domains in which we study cross-temporal IR. This paper has focused on book search. But techniques for handling language change have a role to play in other types of IR. In particular, social media such as microblogs see rapid shifts in discursive conventions [7]. While the temporal dynamics of relevance (e.g. [21]) have seen a good deal of attention recently, the problem of temporality as an invitation for vocabulary mismatch deserves increased scrutiny.

10. CONCLUSION

The growing opportunity for digitized book repositories to impact peoples' use of information suggests that language evolution will be an increasingly important challenge for modern IR. Without prompting, all four of the Medieval Studies Ph.D. candidates who performed our relevance judgments said that they wished that a system capable of cross-temporal search were available to them.

In light of this change, this paper proposed techniques for handling vocabulary mismatch due to temporal shifts in language. From a conceptual standpoint, our contribution entails a generic way of considering the use of texts (dictionaries, corpora, relevant documents) to inform CTIR. More pragmatically, the paper's main contribution is a novel feedback technique that combines several types of evidence to improve cross-temporal retrieval.

11. ACKNOWLEDGEMENTS

This work was supported by a Google Digital Humanities Award and NSF Grant # 1217279. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the funders. The author thanks Jon Orwant for his help in the early stages of this research.

12. REFERENCES

[1] Lisa Ballesteros and W. Bruce Croft. Dictionary methods for cross-lingual information retrieval. In *Proceedings of the 7th International Conference on Database and Expert Systems Applications*, DEXA '96, pp. 791–801, London, UK, 1996. Springer-Verlag.

[2] Adam Berger and John Lafferty. Information retrieval as statistical translation. In *SIGIR '99: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 222–229, New York, NY, USA, 1999. ACM.

[3] Dan Cohen. Is google good for history?, January 7, 2010. http://hdl.handle.net/1920/6101.

[4] Kareem Darwish and Douglas W. Oard. Probabilistic structured query methods. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, pp. 338–344, New York, NY, USA, 2003. ACM.

[5] F.M.G. de Jong, H. Rode, and D. Hiemstra. Temporal language models for the disclosure of historical text. In *Humanities, Computers and Cultural Heritage: Proceedings of the XVIth Intl. Conf. of the Assn. for History and Computing (AHC 2005)*, pages 161–168, Amsterdam,, September 2005. Royal Netherlands Academy of Arts and Sciences.

[6] Michael Droettboom. Correcting broken characters in the recognition of historical printed documents. In *Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '03, pages 364–366, Washington, DC, USA, 2003. IEEE Computer Society.

[7] Miles Efron. Information search and retrieval in microblogs. *Journal of the American Society for Information Science and Technology*, 62(6):996–1008, 2011.

[8] S. Harding, W. Croft, and C. Weir. Probabilistic retrieval of OCR degraded text using n-grams. In Carol Peters and Costantino Thanos, editors, *Research and Advanced Technology for Digital Libraries*, Volume 1324 of *Lecture Notes in Computer Science*, pages 345–359. Springer Berlin / Heidelberg, 1997.

[9] Daqing He and Dan Wu. Enhancing query translation with relevance feedback in translingual information retrieval. *Inf. Proc. and Mgmt.*, 47(1):1 – 17, 2011.

[10] Daniel Jurafsky and James H. Martin. *Speech and Language Processing, 2nd Edition.* Prentice Hall, New York, 2008.

[11] Nattiya Kanhabua and Kjetil Nørvåg. Improving temporal language models for determining time of non-timestamped documents. In *Proceedings of the 12th European Conference on Research and Advanced Technology for Digital Libraries*, ECDL '08, pages 358–370, Berlin, Heidelberg, 2008. Springer-Verlag.

[12] Maryam Karimzadehgan and ChengXiang Zhai. Estimation of statistical translation models based on mutual information for ad hoc information retrieval. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 323–330, New York, NY, USA, 2010. ACM.

[13] Gabriella Kazai and Antoine Doucet. Overview of the INEX 2007 Book Search Track (Booksearch'07). In *INEX*, pages 148–161, 2007.

[14] Gabriella Kazai, Carsten Eickhoff, and Peter Brusilovsky. Report on Booksonline'11: 4th Workshop on Online Books, Complementary Social Media, and Crowdsourcing. *SIGIR Forum*, 46(1):43–50, 2012.

[15] Marijn Koolen, Jaap Kamps, and Gabriella Kazai. Social book search: comparing topical relevance judgements and book suggestions for evaluation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge management*, pages 185–194, 2012.

[16] Marijn Koolen, Gabriella Kazai, Jaap Kamps, Michael Preminger, Antoine Doucet, and Monica Landoni. Overview of the INEX 2012 Social Book Search Track. In *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[17] Abhimanu Kumar, Matthew Lease, and Jason Baldridge. Supervised language modeling for temporal resolution of texts. In *Proceedings of the 20th ACM International Conference on Information and Knowledge management*, CIKM '11, pages 2069–2072, New York, NY, USA, 2011. ACM.

[18] Adenike M. Lam-Adesina and Gareth J. F. Jones. Examining and improving the effectiveness of relevance feedback for retrieval of scanned text documents. *Inf. Proc. Mgmt.*, 42(3):633–649, 2006.

[19] Victor Lavrenko and W. Bruce Croft. Relevance based language models. In *SIGIR '01: Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 120–127, New York, NY, USA, 2001. ACM.

[20] Gina-Anne Levow, Douglas W. Oard, and Philip Resnik. Dictionary-based techniques for cross-language information retrieval. *Inf. Proc. Mgmt.*, 41(3):523–547, 2005.

[21] Xiaoyan Li and W. Bruce Croft. Time-based language models. In *CIKM '03: Proceedings of the 12th International Conference on Information and Knowledge Management*, pages 469–475, New York, NY, USA, 2003. ACM.

[22] Donald Metzler and W. Bruce Croft. A markov random field model for term dependencies. In *SIGIR '05: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 472–479, New York, NY, USA, 2005. ACM.

[23] Donald Metzler and W. Bruce Croft. Latent concept expansion using markov random fields. In *SIGIR '07: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 311–318, New York, NY, USA, 2007. ACM.

[24] Donald Metzler, Victor Lavrenko, and W. Bruce Croft. Formal multiple-bernoulli models for language modeling. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 540–541, New York, NY, USA, 2004. ACM.

[25] Jian-Yun Nie, Michel Simard, Pierre Isabelle, and Richard Durand. Cross-language information retrieval based on parallel texts and automatic mining of parallel texts from the web. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 74–81, New York, NY, USA, 1999. ACM.

[26] J. Oncina and M. Sebban. Learning stochastic edit distance: Application in handwritten character recognition. *Pattern Recognition*, 39(9):1575–1587, 2006.

[27] Marc Parry. The humanities go google. *The Chronicle of Higher Education*, May 2010. http://chronicle.com/article/The-Humanities-Go-Google/65713/.

[28] Matjaz Perc. Evolution of the most common English words and phrases over the centuries. *Journal of the Royal Society: Interface*, 1:3323–3328, 2010.

[29] Eric Sven Ristad and Peter N. Yianilos. Learning string edit distance. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 20(5):522–532, May 1998.

[30] Edward Sapir. *Language: And Introduction to the Study of Speech.* Harcourt, Brace, New York, 1921.

[31] Trevor Strohman, Donald Metzler, Howard Turtle, and W. Bruce Croft. Indri: a language-model based search engine for complex queries. Technical report, in *Proceedings of the International Conference on Intelligent Analysis*, 2005.

[32] J. R. R. Tolkien and E. V. Gordon, editors. *Sir Gawain and the Green Knight.* Clarendon Press, Oxford, 1967.

[33] Justin Zobel and Philip Dart. Phonetic string matching: lessons from information retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 166–172, New York, NY, USA, 1996. ACM.