On Foreign Name Search

Jason Soo[1] and Ophir Frieder[2]

[1] Information Retrieval Laboratory
Illinois Institute of Technology
[2] Department of Computer Science
Georgetown University
soo@ir.iit.edu, ophir@cs.georgetown.edu

Abstract. We address foreign name search in a highly diverse user community. User sophistication ranges from highly experienced archivists to apprehensive users who shy away from technology; apprehensive users dominate system use. Thus, all system interfaces must assume minimal dependency on the user.
Our foreign names search approach, called Segments, is language independent; thus, there is no need to determine the language of origin from the diverse candidate set of thirteen languages. We compare Segments against traditional n-gram and Soundex based solutions. Actual and synthetic queries are used to search a names data set resident in the United States Holocaust Memorial Museum. We also search a subset of the 1990 United States Census Bureau Surnames data set to evaluate the performance of Segments on a predominately language specific (English) collection. Our results demonstrate statistically significant performance gains over both traditional approaches. The described approach supports search efforts at the United States Holocaust Memorial Museum.

1 Introduction

Name identification significantly impacts accuracy in the general search case; however, in historical document search, their identification is paramount. Complicating name search is the variance of accepted spellings for the same sounding name, for example Laurence and Lawrence. To circumvent spelling issues, phonetic search techniques are often used [**?**]. Common phonetic techniques are based on Soundex; JewishGen [**?**] uses the Daitch-Mokotoff (D-M) Soundex variant [**?**], a de facto standard by Jewish genealogical organizations.

A difficulty with using phonetic search stems from the reliance on the user to formulate an approximate sound, and hence spelling, of the foreign name. For example, to an English speaker, *"Roz'ishts' ávarati"* or *"Rozhyshche"* is likely to be difficult to pronounce. Furthermore, in searching name indices from historical documents, particularly for personal-data related applications such as JewishGen genealogy [**?**] or Yizkor Books [**?**], often the user knows that the name of interest has an "esto" in it (from a name like: Nové Mesto nad Váhom) but is uncertain about the remainder of the name or even the language that the name is in. This occurs fairly often since a variety of communities existed in each location. For example, for the German speaking community in Czechoslovakia during the 1930's, "Nové Mesto nad Váhom" was called "Neustadt an der Waag", and "Bratislava" was called "Pressburg" by German speakers and "Pozsony" by Hungarian speakers. In the hope that some fragment of the name will match, which obviously is not always the case, n-gram based solutions [?] are deployed.

Earlier efforts [?,?,?] have demonstrated that efficient simple rules can outperform many traditional approaches. Our language-independent name search approach, called SEGMENTS, follows this trend. That is, we search a collection of foreign names by segmenting the input string according to a set of simple rules. The search results obtained using the individual segments are merged, and a confidence for the merged list is derived. If the confidence is insufficient, namely below a predefined threshold, we invoke an n-gram search.

Extracting from a database of names derived from various documents and texts resident at and/or accessed by the United States Holocaust Memorial Museum, we favorably compare the search accuracy of SEGMENTS against traditional n-gram and D-M Soundex based solutions. Actual user queries as well as synthetic queries generated using single and multiple character addition, deletion, replacement, and inversion are used in our evaluation. We also show favorable results using a subset of the 1990 United States Census Bureau collection of surnames [?]. A subset rather than the entire collection was chosen so as to mirror the size of the United States Holocaust Memorial Museum data set used. SEGMENTS is used in support of search efforts for the United States Holocaust Memorial Museum.

2 Yizkor Book Metadata Search: A SEGMENTS Application

The Yizkor Book Metadata Search project, an effort led by the Archives Section of the United States Holocaust Memorial Museum, aims to create an online metadata global directory of Yizkor Books. Briefly, Yizkor Books memorialize life before, during, and after the Holocaust describing everyday events including births, marriages, and deaths. Many texts were written by survivors or their relatives or friends as a tribute to those who perished. Most texts are written using multiple languages. Thirteen languages are used: Czech, Dutch, English, French, German, Hebrew, Hungarian, Lithuanian, Polish, Romanian, Serbo-Croatian, Spanish, and Yiddish. Given the diversity of languages, only a language independent approach is viable.

Yizkor books are scattered globally, but currently, a global directory of these books is unavailable. Some Yizkor Book repositories go as far as to provide download-ready, scanned copies of the books, for example those residing within the New York City Public Library [?]; however, locating some of these digital repositories, particularly the lesser ones, is accomplished mainly by word of mouth.

The preliminary architecture of Yizkor Books metadata search system was initially described in [?]. Here, in Figure **??**, we illustrate the currently deployed architecture. As shown, metadata are generated internally by USHMM staff or fellows, generated externally by a diversity of users including historians, genealogists, librarians, etc. or are downloaded directly from Yizkor Book repositories and sent via the Internet to the metadata search engine. Once collected, they are organized, temporarily housed in a verification repository (not illustrated), and eventually stored in the Yizkor Books metadata repository. To guarantee correctness, a temporary verification repository is used as an intermediary. That is, all metadata are inspected and verified for accuracy by authorized personnel prior to insertion into the metadata repository. Thus, there are always "humans in the loop".

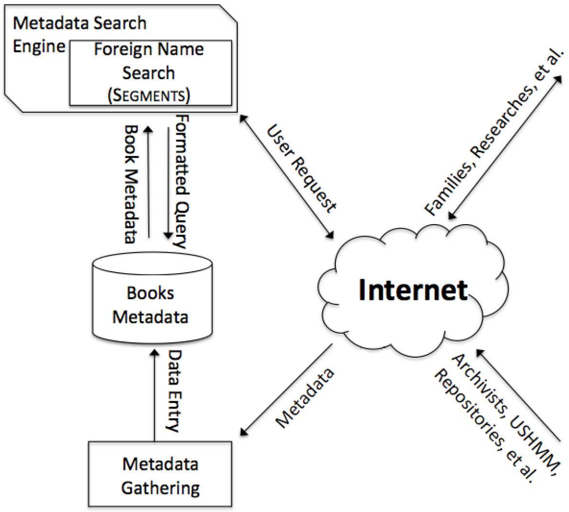

Fig. 1. Yizkor Overview

User queries from a diverse audience are issued and sent likewise via the Internet to the metadata search engine. SEGMENTS is the foreign name search component within the search engine. The queries are translated to the appropriate internal format, issued against the repository, and corresponding metadata are returned. The candidate results are then routed to the requesting party.

3 Algorithm

Our SEGMENTS approach operates as shown in Figure **??**. Initially the user generated query is issued against the name index derived from the collection. No attempt is made to identify the language. If an exact match is found, then the matching name or names and their corresponding information are returned with a confidence of 1. Otherwise, multiple substrings are derived applying simple

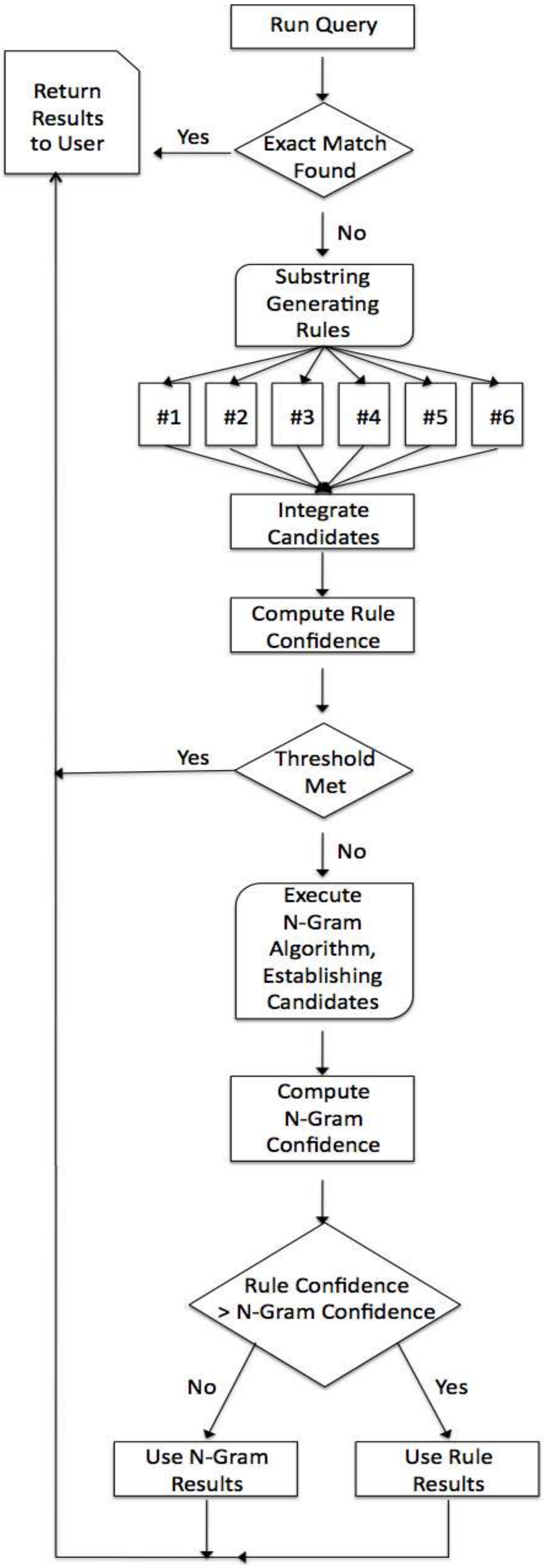

Fig. 2. Query Processing Overview

substring generating rules. The collection is then searched using these derived substrings. Many substring generating rule variations were studied; the deployed system uses:

- Rule 1: Replace first and last characters by a wild card, in succession;
- Rule 2: Replace middle n-characters by a wild card, in succession;
- Rule 3: Replace first half of the string by a wild card;
- Rule 4: Replace second half of the string by a wild card;
- Rule 5: Retain only first and last characters and insert a wild card;
- Rule 6: Retain only first and last two characters and insert a wild card;

Generating Rules 1 and 2 are recursive, and thus, can generate a large number of potential substrings. For Rule 2, typically n=1; however, for substantially long names (greater than 20 characters), n=2. Note that all names are stored in lower-case. To limit the search time, only up to three substrings are generated per rule. Thus, in the deployed implementation, at most ten substrings are generated: three per each of the first two rules and one each for the remaining rules. In future work, we will study in greater detail which SEGMENT rule is most affective and assign. The derived substrings for the search term "Rozhyshche" are illustrated in Table **??**.

Table 1. Actual Query Performance Evaluation

Rule #	Search Candidates		
1	%ozhyshch%	%zhyshc%	%hysh%
2	rozhy%hche	rozh%hche	rozh%che
3	%shche	-	-
4	rozhy%	-	-
5	r%e	-	-
6	ro%he	-	-

For each generated substring, according to its similarity to the desired name, a confidence is determined. Confidence is defined as the summation of all substring matching algorithms which found (voted for) a candidate word over the total number of casted votes. The global confidence of the merged result set is computed. Although multiple generating rules derive some identical candidate strings, this replication was experimentally observed as needed to bolster the confidence. Clearly, our description of the approach is strictly for clarity of presentation; the actual implementation does not replicate the search, rather candidate weights are adjusted accordingly.

If the global confidence fails to meet the needed threshold, an n-gram based solution is deployed. That is, in addition to the segment search already performed, a traditional n-gram search where n=3 is conducted, and a confidence for the n-gram solution is computed. A comparison of the confidence of SEGMENTS and the n-gram solution is made and the option with the higher confidence is selected. Our test data, as well as other related data, are available at http://yizkor.cs.georgetown.edu/collections.

4 Soundex

To illustrate how SEGMENTS avoids problems faced by Soundex and D-M Soundex, a brief overview of these algorithms is necessary.

4.1 Soundex Overview

Soundex masks like-sounding characters by replacing them with integer representations, where said integers map to a set of characters. For example, in Soundex the integer 5 represents either *"m"* or *"n"*. Furthermore, Soundex does not encode the first letter of the given query. Consider the word "Slovakia", which Soundex encodes as "S412" [**?**]. Details of the Soundex algorithm are omitted since only basic knowledge is required to understand the pitfalls.

4.2 Soundex Pitfalls

SEGMENTS addresses multiple known Soundex pitfalls. A subset of these known problems [**?**], which SEGMENTS resolves are:

1. ***Dependence on initial letter.*** If the first letter of the user's query is incorrect, Soundex will never find the correct result [**?**]. SEGMENTS however has 2 rules which will find the correct match.
2. ***Noise intolerance.*** [**?**,**?**,**?**] find that 80%-95% of misspellings within large documents are 1) one character insertions 2) one character deletions 3) one character replacements or 4) adjacent character swapping. Soundex, as demonstrated by [**?**], is unable to reliably resolve such noise. SEGMENTS however has demonstrated its tolerance for noise as shown in Tables **??** and **??**.
3. ***Poor precision.*** One of the strengths of Soundex is the encoding of words as integers representing character groups. This representation however leads to ambiguity and ultimately degrades precision. For example, Soundex encodes the misspelled string "disapont" as "D215". A query would then be run for "D215" which would return: *disband, disbands, disbanded, disbanding, disbandment, disbandments, dispense, dispenses, dispensed, dispensing, dispenser, dispensers, dispensary, dispensaries, dispensable, dispensation, dispensations, deceiving, deceivingly, despondent, despondency, despondently, disobeying, disappoint, disappoints, disappointed, disappointing, disappointedly, disappointingly, disappointment, disappointments, disavowing* [**?**]. Should a query be correctly spelled, Soundex will still return several matches for the same reason. SEGMENTS does not suffer from the same ambiguity due to the voting process of the substring matching algorithms used. Since the above words all have the same encoding, ranking is usually done by frequency [**?**].

4.3 D-M Soundex

D-M Soundex, the Eastern European derivative, adjusts the elements of the character sets for language localization. It also improves upon the Soundex algorithm in the following select ways (localization changes/improvements are omitted)[?]:

1. Encoding of the initial letter. Consider the previous example "Slovakia", which Soundex encodes as "S412". In D-M Soundex "Slovakia" becomes "487500". Notice the extended length of the D-M Soundex encoding, that is the second improvement.
2. The first six (rather than four) significant codes are created. For example, Peters and Peterson have an identical encoding in Soundex ("P362"), but different in D-M Soundex ("739400", "739460").

These changes partially improve some of the downfalls of Soundex (**??**). For example, the first pitfall noted, *Dependence on initial letter*, is clearly solved by the first improvement. The second improvement aids the third pitfall, *Poor precision*, but adds to the time and space complexity. The *noise intolerance* pitfall however is not addressed, and is the root cause of the majority of misspellings. In fact, generally if any word contains more than four (Soundex) or six (D-M Soundex) consonants, all characters thereafter are ignored [?]. Therefore, neither Soundex or D-M Soundex are viable solutions for users, regardless of their knowledge of a language.

5 Evaluation

To evaluate our proposed approach, we randomly selected a subset of roughly 1,000 names from the Jewish Census residing at the United States Holocaust Memorial Museum. Names averaged 8 characters in length, with a median length of 8, a max length of 23, and a standard deviation of 2.8 characters. Using a set of 250 actual queries, we favorably compared the performance of Segments against the popular D-M Soundex approach and a traditional n-gram solution.

Two metrics were used, namely the percentage of names correctly identified and the average rank of those names found. A name was defined as found if it ranked in the top 60 entries (first three screens with 20 names listed per screen). Although we evaluated multiple n-values in the n-gram approach, we present results for only n=3, as it consistently supported the highest percentage found. Average rank rather than MRR is presented as it better illustrates the difference. Undetected entries are ignored in terms of the average rank computation. The statistical t-test was used to verify significance.

In Table **??**, we present our findings using the collected actual 250 queries. Three measures (percentage found, average rank, and common average rank) for each of three approaches (D-M Soundex, 3-Grams, and Segments), when appropriate, are shown. The percentage of correctly identified names is presented in the percentage found column. Since the percentage of correctly identified names is higher when using SEGMENTS rather than 3-grams, a common column is presented to provide direct comparison of SEGMENTS average rank when only considering names also correctly identified by 3-grams. The percentage of correctly identified names is presented in the percentage found column. Both SEGMENTS and 3-Grams statistically significantly ($p<0.01$) outperform D-M Soundex in terms of the percentage of correctly identified names. SEGMENTS likewise statistically significantly ($p<0.01$) outperforms 3-Grams in terms of the percentage of correctly identified names.

Table 2. U.S. Holocaust Memorial Museum Live Query Performance Evaluation

	Percentage Found	**Average Rank**	**Average Rank (Common)**
D-M Soundex	27.56	1.03	N/A
3-Grams	62.17	12.00	N/A
SEGMENTS	78.19	8.26	7.39

In the second column, the average rank (position) of those items found is presented. As shown, the average rank of SEGMENTS is superior to that of 3-Grams. The D-M Soundex average rank is nearly perfect; that is, names correctly identified are almost always positioned first in the rankings. However, this statistic is clearly misleading since roughly only a quarter of the names are correctly identified. What is true, however, is that whenever D-M Soundex recognizes the name, it perfectly identifies it, and this behavior is one possible explanation for the popularity of the D-M Soundex approach. Regardless of its popularity, the poor accuracy provided by D-M Soundex should prohibit its adoption.

All names identified by the 3-Gram approach are likewise identified by SEGMENTS. Given the difference in the percentage detected, clearly SEGMENTS detects additional otherwise unidentified names. To demonstrate how these additional names affect the average rank, we define the common percentage found metric. The common percentage found indicates the rank of those names found by both 3-Grams and SEGMENTS. As shown, the common percentage found is roughly one rank higher than the average rank. This demonstrates that the additionally identified names are, as expected, typically harder to match and increase the average rank.

The above analysis uses 250 actual queries and best represents typical use. However, to systematically evaluate our approach, we repeated the above evaluation, but this time, with an organized set of synthetically generated queries. That is, we randomly added, removed, replaced, and inverted characters in random locations, an approach commonly done to evaluate potential input errors [?]. Deletions were limited so that terms remained at least 4 characters. Three runs were made for each configuration; the averages are reported in Table **??**.

Table 3. U.S. Holocaust Memorial Museum Synthetic Query Performance Evaluation

	D-M Soundex (%)	**N-Gram (%)**	**N-Gram (rank)**	**SEGMENTS (%)**	**SEGMENTS (rank)**
INSERT					
1 char	41.44	94.94	2.55	100	1.71, 1.71
2 char	19.50	91.72	3.45	99.32	2.61, 2.43
3 char	10.67	87.82	4.11	97.52	3.18, 3.02
4 char	6.93	83.85	5.00	95.23	3.87, 3.79
DELETE					
1 char	42.12	93.33	3.45	99.97	2.51, 2.53
2 char	20.41	84.87	4.81	97.96	4.72, 3.95
3 char	11.12	74.68	5.77	92.71	6.42, 4.84
4 char	9.82	70.31	5.95	86.51	7.12, 5.14
REPLACE					
1 char	31.52	92.33	3.27	100	2.15, 2.01
2 char	16.35	80.95	4.49	93.90	4.19, 3.31
3 char	9.29	69.28	5.20	85.61	5.60, 3.87
4 char	5.87	57.81	5.98	75.18	6.85, 4.84
INVERT					
Adj. char	58.01	84.89	4.88	98.00	3.77, 3.00
2 char	17.31	54.59	6.78	71.61	7.38, 5.39
3 char	9.18	42.89	7.40	57.59	8.55, 6.22
4 char	7.09	34.64	8.49	46.76	9.29, 7.26

In Table **??**, the rows represent the various experiments conducted, namely insertion, deletion, replacement, and inversion of 1 to 4 characters. The position of the character(s) in the string is randomly generated using a uniform distribution. In the cases of multiple character inversions, randomly chosen pairs of characters are exchanged sequentially. In the single character inversion case, a single adjacent pair of characters is selected. This special case was chosen so as to match the described errors in [**?**].

As shown, once again, Segments sustains a statistically significant ($p<0.01$) performance improvement over both D-M Soundex and the 3-gram solutions in terms of the percentage of names correctly identified. Likewise, once again, Segments correctly identifies all names detected by the 3-gram approach as well as some additional names. Hence, in the Segments (RANK) column, there are two entries: the first entry represents the average rank for all names identified; the second entry is the common average rank. The Segments approach always sustains a better average ranking when considering only those entries correctly identified by both approaches. In most cases, Segments also continues to sustain a better average ranking overall including those entries not found by the n-gram approach. In the few cases that Segments does not support a higher overall average ranking, the difference is relatively minimal. This occurs when the difference in percentage detection is significant. For all tests conducted, for all names identified by both the 3-Gram and Segments approaches, Segments was statistically significantly (p<0.01) superior.

Table 4. 1990 U.S. Census Bureau Synthetic Query Performance Evaluation

	Soundex (%)	**D-M Soundex (%)**	**N-Gram (%)**	**N-Gram (rank)**	**SEGMENTS (%)**	**SEGMENTS (rank)**
INSERT						
1 char	27.89	28.23	95.69	1.65	100	1.71, 1.65
2 char	8.28	7.76	87.76	2.36	98.02	2.90, 2.36
3 char	2.67	2.29	79.11	3.14	94.73	3.10, 3.13
4 char	1.23	0.48	70.36	3.68	88.79	3.31, 3.68
DELETE						
1 char	45.28	41.87	95.90	2.08	100	2.05, 2.08
2 char	18.49	16.02	80.72	3.47	98.00	5.03, 3.47
3 char	7.11	5.32	62.79	4.14	90.18	7.40, 4.14
4 char	2.48	1.03	48.14	3.96	77.47	8.93, 3.96
REPLACE						
1 char	23.87	23.84	88.60	2.25	99.97	2.46, 2.25
2 char	8.75	8.54	64.75	3.26	84.79	6.02, 3.26
3 char	3.95	3.50	45.72	4.09	67.19	8.07, 4.08
4 char	1.88	1.95	29.23	4.86	51.23	10.75, 4.86
INVERT						
Adj. char	58.34	49.22	72.43	3.28	93.86	5.90, 3.28
2 char	21.65	18.52	31.39	4.16	48.00	10.45, 4.16
3 char	14.25	11.41	27.02	4.48	41.34	11.13, 4.48
4 char	11.96	10.37	21.02	4.74	32.67	11.44, 4.73

Similar evaluation was performed using a subset of the 1990 United States Census Bureau Surnames data set [**?**]. Based on provided statistics, we tested the 1,000 most frequent surnames. A subset was chosen so as to mirror the size of the United States Holocaust Memorial Museum data set used. Names have a mean and median of 6 characters in length, a max of 11, and a standard deviation of 2.4 characters. We, once again, synthetically altered all 1,000 names to generate queries, as previously described. We justify replacing user query logs with machine altered queries on the grounds that given a user who is proficient in a particular language, their queries have a higher probability to contain typos rather than true syntactical errors. As such, random manipulation of query terms results in near real-world examples. Furthermore, such input error testing is commonly done [**?**]. The results are shown in Table **??**. As seen, the relative performance of these algorithms are similar to those obtained using the United States Holocaust Memorial Museum data set. Note Soundex and D-M Soundex have similar performance; hence, all remarks pertaining to D-M Soundex apply to Soundex.

Thus, for both actual queries and for systematically generated synthetic queries, Segments supports a statistically significant ($p<0.01$) performance improvement over both D-M Soundex and a traditional 3-gram solution.

6 Conclusion

To support foreign name identification in an environment in which users recall only distorted portions of desired names, we developed a language-independent, fusion-based, segment-oriented, n-gram supported, search system called Segments. Initially, Segments searches a name index for an exact match. If a name or names are found, they are returned to the user with a confidence of 1. Otherwise, a set of candidate substrings are generated using a set of simple parsing rules. These generated substrings are searched as candidate queries against the name index, and all partially matching names are returned. A confidence for each partial match is computed, and a global confidence for all derived potential result names is likewise computed. The global confidence is compared against a pre-established threshold, and if this threshold is met, a ranked list of derived name candidates, along with the individual confidence of each candidate, is returned to the user. Name candidates are ranked according to their confidence. If, however, the global confidence fails to exceed the pre-established threshold, a traditional n-gram solution is run to derive an additional set of potential result candidates. A global confidence is similarly computed for these candidates. The global confidences for both approaches are compared, and the results corresponding to the higher of the two confidences are returned to the user.

We evaluated our approach using a Jewish Census data set resident at the United States Holocaust Museum and using the 1990 Surnames Census data set from the United States Census Bureau. For our Jewish Census evaluation, to determine expected "real-world" performance, we collected user queries and used them as our initial query test set. As user queries, however, do not necessarily systematically evaluate the approaches under consideration, we likewise created a synthetic query mix derived based on the prior art. That is, actual queries were used to access realistic typical behavior; synthetic queries were used to systematically "stress test" the search system. Our results demonstrate the significantly higher accuracy of our approach as compared to both the D-M Soundex approach presently used initial the JewishGen genealogy search and a traditional n-gram approach using a variety of n values.

Our approach is in current use to enhance search functionally for the United States Holocaust Memorial Museum Yizkor Books effort.

References

1. A. Beider and S. Morse, *Beider-Morse Phonetic Matching: An Alternative to Soundex with Fewer False Hits*, Avotaynu: the International Review of Jewish Genealogy, Summer 2008.

2. M. Aljlayl and O. Frieder, *On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach*, ACM Eleventh Conference on Information and Knowledge Management (CIKM), Washington, DC, November 2002.
3. M. Amir, *From Memorials to Invaluable Historical Documentation: Using Yizkor Books as Resource for Studying A Vanished World*, Annual Convention of the Association of Jewish Libraries, La Jolla, California, June 2001.
4. S. Aqeel, S. Beitzel, E. Jensen, D. Grossman and O. Frieder, *On the Development of Name Search Techniques for Arabic*, Journal of the American Society of Information Science and Technology, 57(6), April 2006.
5. F. Damerau, *A technique for computer detection and correction of spelling errors*, Communications of the ACM, 7(3), pp171-176, March 1964.
6. F. Guy and D. Oard, *The TREC-2001 Cross-Language Information Retrieval Track: Searching Arabic using English, French or Arabic Queries*, NIST TREC, Gaithersburg, Maryland, November 2001
7. JewishGen, September 1, 2009. http://jewishgen.org.
8. C. Manning, P. Raghavan, H. Schutze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.
9. R. Mitton, *Spellchecking by Computers*, Journal of the Simplified Spelling Society, J20, 1996.
10. G. Mokotoff, *Soundexing and Genealogy*, 2007. September 1, 2009. http://www.avotaynu.com/soundex.html.
11. New York Public Library Yizkor Books, September 1, 2009. http://www.nypl.org/research/chss/jws/yizkorbookonline.cfm.
12. F. Patman and L. Shaefer, *Is Soundex Good Enough for You? On the Hidden Risks of Soundex-Based Name Searching*, Language Analysis Systems, Inc., Herndon, VA, 2003.
13. J. Pollock, and A. Zamora, *Automatic spelling correction in scientific and scholarly text*, Communications of the ACM, 27(4), April 1984.
14. C. Snae and M. Bruckner. *Novel Phonetic Name Matching Algorithm with a Statistical Ontology for Analysing Names Given in Accordance with Thai Astrology*, Issues in Informing Science and Information Technology, 2009.
15. J. Soo, R. Cathey, O. Frieder, M. Amir, and G. Frieder, *Yizkor Books: A Voice for the Silent Past*, ACM Seventeenth Conference on Information and Knowledge Management (CIKM), Napa Valley, California, October 2008.
16. United States Census Bureau 1990 Surnames, September 1, 2009. http://www.census.gov/genealogy/names/dist.all.last.
17. J. Zobel and P. Dart, *Phonetic String Matching: Lessons from Information Retrieval*, ACM Nineteenth Conference on Research and Development in Information Retrieval (SIGIR), Zurich, Switzerland, August 1996.