LAMP-TR-110 November, 2003
CS-TR-4549
UMIACS-TR-2003-117

A SIMILARITY-BASED APPROACH AND EVALUATION METHODOLOGY FOR REDUCTION OF DRUG NAME CONFUSION

Grzegorz Kondrak, Bonnie J. Dorr

Department of Computer Science, University of Alberta, Canada and
Institute for Advanced Computer Studies, University of Maryland
kondrak@cs.ualberta.ca, bonnie@umiacs.umd.edu

Abstract

This paper facilitates the task of mitigating medical errors due to the confusion of look-alike and sound-alike drug names. Detection of potential confusion is based on both feature-based phonetic comparison (for sound-alike drug names) and orthographic similarity (for look-alike drug names). We present a new recall-based evaluation methodology for determining the effectiveness of different similarity measures on drug names. Using this methodology, we show that a new orthographic measure called BI-SIM outperforms other commonly used measures of similarity on a set containing both look-alike and sound-alike pairs, In addition, we demonstrate that the feature-based phonetic approach outperforms other standard approaches on a test set containing solely look-alike confusion pairs. However, an approach that combines several different approaches achieves the best results on both test sets.

Keywords: phonological and orthographic similarity, drugname confusion

The support for the LAMP Technical Report Series and the partial support of this research by the Department of Defense under contract MDA9049-C6-1250, NSF ITR Grant IIS-0326553, and NSF CISE Research Infrastructure Award EIA0130422 is gratefully acknowledged.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2003		**00-00-2003 to 00-00-2003**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
A Similarity-Based Approach and Evaluation Methodology for Reduction of Drug Name Confusion	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **University of Maryland,Institute for Advanced Computer Studies,College Park,MD,20742**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**		**16**	

1 Introduction

Many hundreds of drugs have names that either look or sound so much alike that doctors, nurses and pharmacists can get them confused, dispensing the wrong one in errors that can injure or even kill patients. In the United States alone, an estimated 1.3 million people are injured each year from medication errors, such as administering the wrong dose or the wrong drug [Lazarou *et al.*, 1998]. For example, a patient needed an injection of Narcan but instead got the drug Norcuron and went into cardiac arrest. The U.S. Food and Drug Administration has sought to mitigate this threat by ensuring that proposed drug names that are too similar to previously existing drug names are not approved [Meadows, 2003]. This paper proposes an algorithmic approach to addressing this need.

Our approach involves the application of two new methods to the problem of detecting potential drug-name confusion, one based on phonetic similarity ("sound-alike") and the other based on orthographic similarity ("look-alike"). We demonstrate that a combined approach provides the best result on a test set that contains both look-alike and sound-alike confusion pairs. By providing computerized techniques for detecting similarity, we induce a reliable, reproducible, automatically evaluable approach, as advocated by well-known researchers in the field [Lambert, 1997].

For the detection of sound-alike confusion pairs, we apply a string-matching algorithm that involves the phonetic transcription of two strings followed by the application of a feature-based phonetic comparison. The approach proposed is based on the ALINE cognate matching algorithm [Kondrak, 2001] which calculates the similarity between word pairs based on their phonetic features.

Consider the example of *Xanax* vs. *Zantac*—two brand names that the *Physicians' Desk Reference* (PDR) warns may be "mistaken for each other ... lead[ing] to serious medication errors" [24th Ed., 2003]. The phonetic transcription of the two names, [zæn^ks] and [zænt^k], reveals their sound-alike similarity that is not apparent in their orthographic form. ALINE assigns a similarity score to the pair of phonetic strings that is based on the optimal alignment of both identical and similar phonemes. The similarity between individual phonemes is evaluated on the basis of their decomposition into elementary phonetic features.

	Distance	Similarity
Orthographic	EDIT	DICE, LCSR
Phonetic	SOUNDEX	ALINE

Table 1: Classification of Distance/Similarity Measures

For the detection of look-alike confusion pairs, we propose a new measure of orthographic similarity, called BI-SIM, that combines the advantages of several known approaches. We show that this new measure performs better than other commonly used measures of similarity.

Validation of the phonetic and orthographic approaches is based on measuring the recall against an on-line gold standard. We have conducted experiments on a U.S. pharmacopeial gold standard which contains 399 true confusion pairs involving 582 unique drug names. Several drug-name matching approaches were compared using a new recall-based evaluation methodology for determining the effectiveness of different similarity measures. We demonstrate that on a set containing both look-alike and sound-alike pairs, BI-SIM achieves the best results, while on the test set containing solely look-alike confusion pairs, ALINE-based similarity matching outperforms other approaches. However, an approach that combines several different approaches attains even higher recall levels on both test sets.

2 Background

Drug-name matching refers to the process of string matching to rank similarity between drug names. There are two classes of string matching: orthographic and phonetic. For each of these, there are two methods of matching: distance and similarity. Our hypothesis is that, if two drug names are confusable, the distance between them will be small and the similarity between them will be large. We refer to this below as the *drug-name confusion similarity/distance* (DCSD) hypothesis.

Some examples of orthographic and phonetic algorithms for both distance- and similarity-based approaches are shown in Table 1.

String-edit distance [Wagner and Fischer, 1974]—also known as Levenshtein distance—counts up the number of steps it takes to transform one string into another, where the cost of

Code	Letters	Code	Letters
0	a,e,h,i,o,u,w,y	1	b,f,p,v
2	c,g,j,k,q,s,x,z	3	d,t
4	l	5	m,n
6	r		

Table 2: Soundex Character Conversion

substitution is same as the cost of insertion or deletion.[1] For example, the string-edit distance between *Zantac* and *Contac* is 2, whereas the distance between *Zantac* and *Xanax* is 3. A normalized version of edit distance is calculated by dividing the total edit cost by the length of the longer string. Thus, the distance between *Zantac* and *Contac* is $\frac{2}{6} = .33$, whereas the distance between *Zantac* and *Xanax* is $\frac{3}{6} = .5$. In this case, the distance scores are not consistent with the DCSD hypothesis because the pair *Zantac* and *Xanax* is more likely to be confused than *Zantac* and *Contac*.

Two additional forms of orthographic matching are LCSR [Melamed, 1999] and DICE [Adamson and Boreham, 1974], both of which provide a measure of similarity rather than a measure of distance. The LCSR approach divides the length of the longest common subsequence by the length of the longer string. For example, the LCSR between *Zantac* and *Contac* is $\frac{4}{6} = .67$, whereas the LCSR between *Zantac* and *Xanax* is $\frac{3}{6} = .5$. The DICE approach counts the number of shared bigrams prior to dividing by the total number of bigrams in each string.[2] For example, the DICE similarity between za,an,nt,ta,ac and co,on,nt,ta,ac is $\frac{2 \cdot 3}{5+5} = \frac{6}{10} = .6$, whereas the DICE similarity between za,an,nt,ta,ac and xa,an,na,ax is $\frac{2 \cdot 1}{5+4} = \frac{2}{9} = .22$. Again—for both scoring algorithms—the results are not consistent with the DCSD hypothesis: *Zantac* and *Xanax* should be "more similar" than *Zantac* and *Contac*.

Phonetic alternatives to drug name matching have also been proposed. One of the most common approaches is Soundex [Hall and Dowling, 1980], which transforms all but the first letter to numeric codes (see Table 2) and after removing zeroes truncates the resulting string to 4 characters. This approach is able to detect certain sound similarities, while missing others. For

[1] There is also a variant of Levenshtein where the cost of substitution is twice the cost of insert or delete; after normalization, this version is equivalent to the complement of the Longest Common Subsequence Ratio (i.e., 1 minus the LCSR).

[2] DICE is described here with bigrams, but it could be applied with arbitrarily large n-grams.

example, the approach is capable of finding a match between the two sound-alike words *king* and *khyngge* (k520,k520), but it is unable to detect a match between *knight* and *night*. Even worse, Soundex matches radically different sounding words such as *pulpit* and *phlebotomy* (p413,p413). For the purposes of comparison, we implemented a Soundex-based similarity measure that returns the edit distance between the corresponding codes. For example, the distance between the Soundex renderings of *Zantac* (z532) and *Xanax* (x520) is 3, while the distance between the Soundex renderings of *Zantac* (z532) and *Contac* (c532) is 1 (which is inconsistent with the DCSD hypothesis).

The deficiency of approaches like Soundex has long been recognized. In a cogent review by [Zobel and Dart, 1995], extensive experiments were executed using a variety of different approaches and their combinations. This study showed that orthographic approaches (like string-edit) were superior to their phonetic alternatives in tasks involving string matching, based on the IR metrics of precision and recall. However, Zobel and Dart's study pre-dates the current state-of-the-art in phonetic coding approaches, most notably that of ALINE (to be described next). Moreover, Soundex was not designed to be applied to the task inherent in Zobel and Dart's tests, i.e. orthographic matching. Rather, these approaches were designed to detect similar-sounding names, which implies that a different, more phonetically-oriented test is needed in order to evaluate those approaches appropriately.

The remainder of this paper addresses some of the deficiencies described above and also proposes a new evaluation methodology for determining the effectiveness of different similarity measures and their combinations.

3 Measuring Phonetic Similarity with ALINE

An alternative to the orthographic similarity/distance and phonetic distance approaches discussed above is a phonetic similarity approach called ALINE [Kondrak, 2000], which uses phonetic features to compare two words by their sounds. ALINE is designed to address some of the issues with algorithms like Soundex: (1) it uses the entire string instead of truncating word to four characters; (2) it involves vowels in the matching process instead of dropping them out; and (3) it uses decomposable features instead of numbers.

		DICE	LCSR	SOUNDEX	ALINE
zantac	xanax	0.222	0.500	0.250	0.733
zantac	contac	0.600	0.667	0.750	0.567
xanax	contac	0.000	0.333	0.250	0.367

Table 3: Comparison of Scores for *Zantac*, *Xanax* and *Contac*.

In this approach, phonetic similarity is established between two words as a by-product of finding an optimal match between their corresponding phonetic features. The similarity value is normalized by the length of the longer word. For example, the *Zantac*/*Xanax* pair is assigned a matching value of 0.733. This is a higher value than that of *Zantac*/*Contac* (0.567) and *Xanax*/*Contac* (0.367). In particular, the words *Xanax* and *Zantac* are considered similar (even if not an exact match) in their initial sound because the word-initial letter "x" of *Xanax* is mapped into the same consonant as the letter "z" (voiced, alveolar, fricative).[3]

There are two fundamental components of ALINE: (1) a similarity function that uses linguistic feature analysis measurements based on *salience,* e.g., the features *Alveolar* and *Stop* are more salient than *Voice*; and (2) a method for choosing optimal alignment that is based on a weighted multi-feature analysis. The approach is designed to align phonetic sequences for many different computational-linguistics applications and, in fact, was initially designed to identify cognates in vocabularies of related languages (e.g. *colour* and *couleur*) [Kondrak, 2000]. Phonetic features are associated with weights that can be fine-tuned for a specific application. The approach uses a (quadratic) dynamic programming algorithm for finding the optimal alignment.[4]

In our initial investigation, we found ALINE to be consistent with the DCSD in many cases where several other algorithms were not. Some examples involving the *Zantac*/*Contac*/*Xanax* pairs given earlier are shown in Table 3.

Prior to running ALINE to test its capability of detecting drug-name confusion, the algorithm requires fine tuning so that it covers drug-name data more adequately; parameters have

[3]The term consonant refers roughly to non-vowel; alveolar refers to the sound being produced just behind teeth; fricative refers to friction during sound production; and voiced refers to sound being produced by vocal cords.

[4]Our choice of ALINE over alternative phonetic-similarity approaches (e.g., [Covington, 1996] and [Somers, 1998]) is due to its favorable comparison over such approaches in a rigorous study [Kondrak, 2000].

default settings for the cognate matching task, but these settings are not appropriate for drug-name matching. Parameter tuning refers to calculation of weights for a particular task, in this case drug-name matching, and then running a hill-climbing search against a gold standard. The parameters that are tuned for the drug-name task are: (i) maximum score; (ii) insertion/deletion penalty; and (iii) vowel penalty.

4 BI-SIM — New Measure of Orthographic Similarity

An analysis of the similarity values computed by commonly used similarity measures reveals their weaknesses. In spite of its popularity, the DICE coefficient is an example of a measure that is demonstrably inappropriate for measuring word similarity. Because it is based exclusively on bigrams, it often fails to discover any similarity between words that look very much alike. For example, it returns zero on the pair *Verelan/Virilon*. In addition, it violates a desirable requirement of any similarity measure—consistent with the DCSD hypothesis—that the maximum similarity of 1 should only result when comparing identical words. In particular, non-identical pairs[5] like *Xanex/Nexan*—where *all* bigrams are shared—are assigned a similarity value of 1. Moreover, it sometimes associates bigrams that occur in radically different word positions, as in the pair *V***ol***taren/Tramad***ol**. Finally, the initial segment, which is arguably the most important in determining drug-name confusability,[6] is actually given *lower* weight than other segments because it participates in only one bigram.

We have observed that the LCSR—which combines unigram resolution with the no-crossing-links constraint[7]—is more appropriate for identifying potential drug-name confusability because it does not rely on (frequently imprecise) bigram matching used in DICE. On the other hand, LCSR is weak in its tendency to posit non-intuitive links, such as the ones between segments in *Ben***adr***yl/C***a***r***d***u***r***a*. The fact that it returns the same value for **Am***aryl/***Am***ikin* and for **A***mary***l/Al***toce* can be attributed to lack of context sensitivity characteristic for unigram-based measures.

[5] This observation is due to [Ukkonen, 1992].

[6] 74.2% of the confusable pairs in the pharmacopeial gold standard (Section 6) have identical initial segments.

[7] The no-crossing-links constraint states that the matched n-grams must form a subsequence of both of the compared strings. In DICE, the order of n-grams is insignificant.

BI-SIM (X, Y)

$m \leftarrow$ length(X)
$n \leftarrow$ length(Y)
$x_0 \leftarrow x'_1$
$y_0 \leftarrow y'_1$
for $i \leftarrow 0$ **to** m **do**
    $F[i, 0] \leftarrow 0$
for $j \leftarrow 1$ **to** n **do**
    $F[0, j] \leftarrow 0$
for $i \leftarrow 1$ **to** m **do**
    for $j \leftarrow 1$ **to** n **do**
        $F[i, j] \leftarrow max($
            $F[i-1, j-1] + id(x_{i-1}, y_{j-1}) + id(x_i, y_j),$
            $F[i-1, j],$
            $F[i, j-1])$
return $F[m, n]/(2 \times max(m, n))$

$$id(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases}$$

Figure 1: BI-SIM algorithm for Computing Similarity of Strings X and Y.

An analysis of the reasons behind unsatisfactory performance of commonly used measures led us to propose a new measure of orthographic similarity, called BI-SIM. This measure aims at combining the advantages of the context inherent in bigrams, the precision of unigrams, and the strength of the no-crossing-links constraint. BI-SIM identifies the longest subsequence of both identical and similar bigrams and normalizes its length by the length of the longer string.

Figure 1 contains pseudo-code that can be used to compute BI-SIM. The pseudo-code exhibits strong similarity to the well-known dynamic-programming algorithm for computing longest common subsequence. The difference lies in the fact that the subsequence is composed of bigrams rather than unigrams, and that the bigrams are weighted according to their similarity. In order to preserve the salience of the initial segment, a corresponding extra symbol is appended to the beginning of each string. The returned value of BI-SIM always falls in the interval $[0, 1]$; in particular, it returns 1 if and only if the strings are identical, and 0 if and only if the strings have no segments in common. Table 4 compares the values computed by BI-SIM for some word pairs with the values returned by DICE and LCSR.

		DICE	LCSR	BI-SIM
ara	ala	0.000	0.677	0.667
atara	arata	1.000	0.600	0.600
amaryl	amikin	0.100	0.333	0.417
amaryl	altoce	0.000	0.333	0.250

Table 4: Comparison of Scores of DICE, LCSR, and BI-SIM for Some Word Pairs.

BI-SIM, as it is defined here, distinguishes three levels of bigram similarity: 2 for identical bigrams, 1 for partly identical bigrams, and 0 for completely distinct bigrams. In principle, the scale could be further refined to include more levels of similarity. For example, bigrams that are frequently confused because of their typographic or cursive shape, such as *en/im*, could be assigned a similarity value that corresponds to the frequency of their confusions. The measure can also be generalized to include arbitrary n-grams.

5 Evaluation Methodology

We designed a new method of evaluating the accuracy of a measure. For each drug name, we sort all the other drug names in the test set in the order of decreasing value of similarity. We calculate the *recall* by dividing the number of true positives among the top k names by the total number of true positives for this particular drug name, i.e., the fraction of the confusable names that are discovered by taking the top k similar names. At the end we apply an information-retrieval technique called *macro-averaging* [Salton, 1971] which takes an average of the recall values across all of the drug names in the test set.[8]

Because there is a trade-off between recall and the k threshold, it is important to measure the recall at different values of k. Table 5 shows the top 10 names that are most similar to *accupril* according to the DICE similarity measure. A '+'/'–' mark indicates whether the pair is a true confusion pair. The pairs are listed in rank order, according to the score assigned by the indicated algorithm. Names that return the same similarity value are listed in the reverse

[8]We could have also chosen to micro-average the recall values instead by dividing the total number of true positives discovered among the top k candidates by the total number of true positives in the test set. The choice of macro-averaging over micro-averaging does not affect the relative ordering of similarity measures implied by our results.

1.	accupril	0.4286	prinivil	–	0.00
2.	accupril	0.4286	prilosec	–	0.00
3.	accupril	0.4286	ocupress	–	0.00
4.	accupril	0.4286	monopril	+	0.33
5.	accupril	0.4286	bepridil	–	0.33
6.	accupril	0.4286	accutane	+	0.67
7.	accupril	0.4000	lacrilube	–	0.67
8.	accupril	0.4000	enalapril	–	0.67
9.	accupril	0.4000	captopril	–	0.67
10.	accupril	0.3750	lisinopril	–	0.67

Table 5: Top 10 Names that are Most Similar to *Accupril* according to DICE Similarity Measure and the Corresponding Recall Values.

lexicographic order. Since the test set contains three drug names that have been identified as confusable with *Accupril* (*Aciphex*, *Accutane*, and *Monopril*), the recall values are 0.33 for $k = 5$, and for 0.67 for $k = 10$.

6 Experiments and Results

We conducted two experiments with the goal of evaluating the relative accuracy of several measures of similarity in identifying confusable drug names, The first experiment was performed against an on-line gold standard: the *United States Pharmacopeial Convention Quality Review, 2001* (henceforth referred to as the *USP set*). The USP set contains both look-alike and sound-alike confusion pairs. We used 582 unique drug names from this source to combinatorically induce 169,071 possible pairs. Out of these, 399 were true confusion pairs in the gold standard. Table 6 shows the number of drug names according to the number of corresponding confusable names in the test set. The maximum number of true positives was 6, but for the majority of names, only one confusable name is identified in the gold standard. On average, the task was to identify 1.37 true positives among 581 candidate names.

We computed the similarity of each name pair using the following similarity measures: PREFIX, DICE, LCSR, EDIT, SOUNDEX, BI-SIM and ALINE. PREFIX is a baseline-type similarity measure that returns the length of the common prefix divided by the length of the longer string. EDIT refers to the normalized version of edit distance, which consistently out-

true positives	1	2	3	4	5	6
test names	436	93	39	12	1	1

Table 6: Number of Test Names with a Given Count of True Positives in the USP Set.

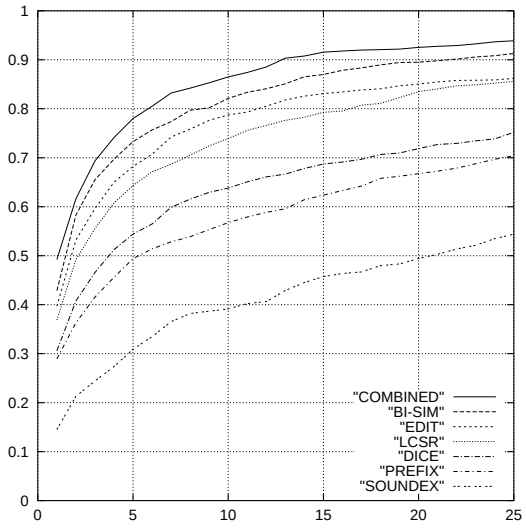

Figure 2: Recall at Various Thresholds for the USP Test Set.

performed raw edit distance in our experiments. COMBINED was calculated by taking the simple average of the values returned by PREFIX, EDIT, BI-SIM, and ALINE.

In order to apply ALINE to the USP data, we have transcribed all names into phonetic symbols. The phonetic transcription of drug names was approximated by applying a simple set of about thirty regular expression rules. (It is likely that a more sophisticated transcription method would result in the improvement of ALINE's performance.) In the first experiment, the parameters of ALINE were not optimized; rather, they were set according to the values specified by [Kondrak, 2001] for a distinct task of cross-language cognate identification.

Figure 2 shows the macro-averaged recall values plotted against k for each of the algorithms on the USP set. The curve that corresponds to ALINE is very close to the curve that corresponds to EDIT, and therefore it has been omitted in order to maintain the clarity of the plot.

We felt that the USP set, which contains both look-alike and sound-alike name pairs, is not a fair test for the phonetic similarity measures, such as ALINE and SOUNDEX. Therefore, we

true positives	1	2	3	4	5	6
test names	15	23	12	14	7	6
true positives	7	8	9	10	11	12
test names	1	2	2	0	1	0

Table 7: Number of Test Names with a Given Count of True Positives in the Sound-alike Set.

conducted a second experiment using a proprietary list of sound-alike drug names. The list contains 276 drug names identified by experts as "names of concern" for 83 "consult" names. None of the "consult" names, and only about 25% of the "names of concern" are encountered in the USP set, which means that there are no true positive pairs shared between the two sets. Table 7 shows the number of drug names according to the number of corresponding confusable names in the test set. The maximum number of true positives was 11, while the average for all names was 3.33.

The measures were applied to calculate the similarity between each of the 83 "consult" names and a list of 2596 drug names. The results are shown in Figure 3. Since the task, which involved identifying, on average, 3.33 true positives among 2596 candidates, was much more challenging, the recall values are lower than in Figure 2. All drug names were first converted into a phonetic form by means of a set of regular expression rules. (We found that phonetic transcription led to a slight improvement in the recall values achieved by the orthographic measures.) The parameters of ALINE used in this experiment were optimized beforehand on the USP set.

7 Discussion

The results described in Section 6 clearly indicate that BI-SIM, the newly proposed measure of similarity, outperforms several currently used measures on the USP test set regardless of the choice of the cutoff parameter k. However, a simple combination of several measures achieves even higher accuracy. On the sound-alike confusion set, ALINE is the top performer. The accuracy achieved by the best measures is impressive. For the combined measure, the average recall on the USP set exceeds 90% with only 13 top candidates considered.

It is important to note that the USP test set has its limitations. The set includes pairs that

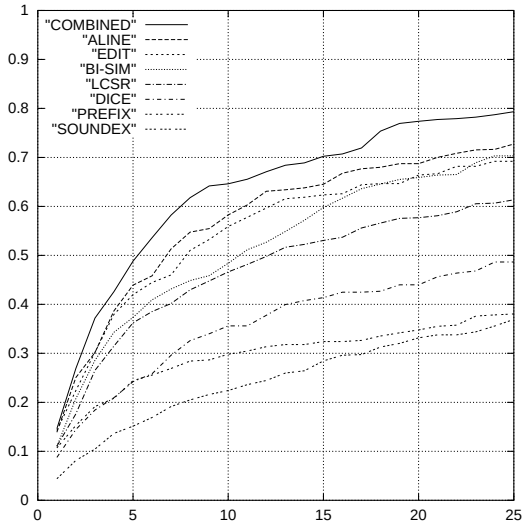

Figure 3: Recall at Various Thresholds for the Sound-alike Test Set.

are considered confusable for other reasons than just phonetic or orthographic similarity, including illegible handwriting, incomplete knowledge of drug names, newly available products, similar packaging or labeling, and incorrect selection of a similar name from a computerized product list. In many cases, the names do not sound or look alike, but when handwritten or communicated verbally, these names have caused or could cause a mix-up. On the other hand, many clearly confusable name pairs are not identified as such (e.g. *Erythromycin/Erythrocin*, *Neosar/Neoral*, *Lorazepam/Flurazepam*, *Erex/Eurax/Urex*, etc.).

All similarity measures have their own strengths and weaknesses. DICE is effective at recognizing pairs such as *Chlorpromazine/Prochlorperazine*, in which a shorter name closely matches parts of the longer name. However, this advantage is offset by its poor performance on similar-sounding names with few shared bigrams (*Nasarel/Nizoral*). LCSR is able to identify pairs in which the common subsequence is interweaved with dissimilar segments, such as *Asparaginase/Pegaspargase*, but fails on similar sounding names in which the overlap of identical segments is minimal (*Luride/Lortab*). ALINE detects phonetic similarity even when it is obscured by the orthography (eg. *Xanax/Zantac*), but it requires phonetic transcription to be performed beforehand.

The idiosyncrasies of individual measures are attenuated when they are combined together,

which may explain the excellent performance of the combined measure. Each measure is focused on a particular facet of string similarity: initial segments in PREFIX, phonetic sound-alike quality in ALINE, common clusters in bigram-based measures, overall transformability in EDIT, etc. For this reason, a synergistic blend of several measures achieves higher accuracy than any of its components.

8 Conclusions and Future Work

We have proposed a new measure of orthographic similarity that is applicable to the problem of detecting drug-name confusion. We have shown that the new measure outperforms several commonly used similarity measures on a publicly available gold standard of confusable drug names. Our results suggest that a linear combination of several measures benefits from the strengths of its components, and is likely to outperform any individual measure. Such a combined approach has the potential to provide the basis of automatic minimization of medication errors.

An area of future work is the development and use of an interface that allows applicants to enter newly proposed drug names. The interface would display a set of scores (one for each potentially confusable drug name) returned by each algorithm, by matching the proposed name against a pre-existing database of currently existing drug names; also included would be a set of scores based on the union of all the approaches. The software would be used by drug companies prior to their submission of a drug name for approval. The applicant would compare the score returned by the algorithms to a pre-determined threshold in order to assess appropriateness of the proposed name.

References

[24th Ed., 2003] PDR 24th Ed. *Physicians' Desk Reference for Nonprescription Drugs and Dietary Supplements*. Thomson PDR, New York, NY, 2003.

[Adamson and Boreham, 1974] George W. Adamson and Jillian Boreham. The use of an association measure based on character structure to identify semantically related pairs of words and document titles. *Information Storage and Retrieval*, 10:253–260, 1974.

[Covington, 1996] Michael A. Covington. An algorithm to align words for historical comparison. *Computational Linguistics*, 22(4):481–496, 1996.

[Hall and Dowling, 1980] Patrick A. V. Hall and Geoff R. Dowling. Approximate string matching. *Computing Surveys*, 12(4):381–402, 1980.

[Kondrak, 2000] Grzegorz Kondrak. A New Algorithm for the Alignment of Phonetic Sequences. In *Proceedings of the First Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 288–295, Seattle, WA, 2000.

[Kondrak, 2001] Grzegorz Kondrak. Identifying Cognates by Phonetic and Semantic Similarity. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics*, Pittsburgh, PA, 2001.

[Lambert, 1997] Bruce L. Lambert. Predicting Look-alike and Sound-alike Medication Errors. *American Journal of Health-System Pharmacy*, 54(10):1161–1171, 1997.

[Lazarou *et al.*, 1998] J. Lazarou, B.H. Pomeranz, and P.N. Corey. Incidence of Adverse Drug Reactions in Hospitalized Patients. *Journal of the American Medical Association*, 279:1200–1205, 1998.

[Meadows, 2003] Michelle Meadows. Strategies to Reduce Medication Errors. *U.S. Food and Drug Administration Consumer Magazine*, May-June, 2003.

[Melamed, 1999] I. Dan Melamed. Bitext maps and alignment via pattern recognition. *Computational Linguistics*, 25(1):107–130, 1999.

[Salton, 1971] Gerard Salton. *The Smart System: Experiments in Automatic Document Processing*. Englewood Cliffs: Prentice Hall, NJ, 1971.

[Somers, 1998] Harold L. Somers. Similarity metrics for aligning children's articulation data. In *Proceedings of COLING-ACL'98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, pages 1227–1231, 1998.

[Ukkonen, 1992] Esko Ukkonen. Approximate string-matching with q-grams and maximal matches. *Theoretical Computer Science*, 92:191–211, 1992.

[Wagner and Fischer, 1974] Robert A. Wagner and Michael J. Fischer. The string-to-string correction problem. *Journal of the ACM*, 21(1):168–173, 1974.

[Zobel and Dart, 1995] Justin Zobel and Philip W. Dart. Finding approximate matches in large lexicons. *Software — Practice and Experience*, 25(3):331–345, 1995.