Advances in Automatic Chemical Spelling Correction

Roger Sayle and Daniel Lowe
NextMove Software
Cambridge, UK
example spelling errors

• Sample misspellings of pyrimidine in US patent grants since the beginning of this year.

<table>
<thead>
<tr>
<th>Incorrect Name</th>
<th>US Patent No.</th>
<th>Issue Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>pryimidine</td>
<td>8093264</td>
<td>10th January 2012</td>
</tr>
<tr>
<td>pyrimidine</td>
<td>8097728</td>
<td>17th January 2012</td>
</tr>
<tr>
<td>pyrmdine</td>
<td>8114996</td>
<td>14th February 2012</td>
</tr>
<tr>
<td>pyrimidne</td>
<td>8129897</td>
<td>6th March 2012</td>
</tr>
<tr>
<td>pyrimidinc</td>
<td>8148339</td>
<td>3rd April 2012</td>
</tr>
<tr>
<td>pyridmidine</td>
<td>8158627</td>
<td>8th May 2012</td>
</tr>
</tbody>
</table>
previous work


The Levenshtein Distance (Levenshtein 1965) is the minimum number of edits (insertions, deletions or substitutions) required to transform one string into another.

The Damerau-Levenshtein Distance is an extension of Levenshtein Distance to include transposition of two adjacent characters.

The distances can be efficiently computed with dynamic programming using the Needleman-Wunsch-Sellers alignment algorithm (bioinformatics).
needleman-wunsch-sellers
needleman-wunsch-sellers
further complications

• Edit operations can have their own specific penalties.
• The latest implementation supports transpositions, to catch spelling mistakes such as “chlorofrom”.
• Some dictionaries to match against have tens of millions of entries, others are infinite.
• The start and end of the input isn’t known in free-text and are assigned on the quality of the match.
• Correct nesting of parenthesis and brackets needs to be enforced as part of the matching process.
• In summary - It’s mind bogglingly complicated.
Nitrogen containing heterocycles as minimal DFA:
- Pyrrole, Pyrazole, Imidazole, Pyrdine, Pyridazine, Pyrimidine, Pyrazine
Example iupac-like grammar

• More generally, still CaffeineFix FSMs can represent formal grammars, i.e. infinite dictionaries.

\[\begin{align*}
\text{alk} & := \text{"meth" | "eth" | "prop" | "but"} \\
\text{parent} & := \text{alk "ane"} \\
\text{subst} & := \text{"bromo" | "chloro" | "fluoro"}
\end{align*}\]

\[\begin{align*}
\text{locant} & := \text{"1" | "2" | "3" | "4" /* any digit */} \\
\text{prefix} & := [ \text{prefix "-" } ] [ \text{loc "-" } ] \text{subst} \\
& \quad | [ \text{prefix } ] \text{subst} \\
\text{name} & := [ \text{prefix [ "-" ] } ] \text{parent}
\end{align*}\]
Iupac-like grammar examples

- methane
- chloroethane
- 2-bromo-propane
- chloro-bromo-methane
- 1-fluoro-2-chloro-ethane
- chlorofluoromethane
- 4-bromomethane
- 1-chloro-1-chloro-1-chloro-1-chloro-methane
Representing grammars as dFAs

Backward edges allow matching an infinite number of words.
current iupac grammar FSM

• As of July 2012, the current CaffeineFix grammar contains nearly 1.2 million edges.

• This grammar covers...
  – 99.20% (232271/234142) names in the NCI00 database.
  – 95.90% (68439/71367) names in the Maybridge catalogue.
  – 95.49% (45996/48167) names in the Keyorganics catalogue.

• These figures are comparable to name-to-structure conversion rates on these names.
SPelling correction examples

- 1H-ben zimidazole → 1H-benzimidazole
- triphenylposhine → triphenylphosphine
- 4- (2-ADAMANTYLCARBAM0YL) -5-TERT-BUTYL-PYRAZOL-1-YL] BENZOIC ACID → 4-(2-adamantylcarbamoyl)-5-tert-butyl-pyrazol-1-yl]benzoic acid
- didec-2-ene → dodec-2-ene
- spiro[2.2]hexane → spiro[2.3]hexane
low cost “frequent” edit ops

- A number of common corrections are so frequent as to be given a lower (free) cost.
  1. Deletion of whitespace.
  2. Deletion of a hyphen (where not anticipated)
  3. Substitution of “l” (lower case el) for “1” (one).
  4. Substitution of “l” (upper case ey) for “l” (el) or “1” (one).
  5. Substitution of “rn” by “m”.
  6. Substitution of “1” (one) by “l” (el).
  7. Substitution of “&phi;” by “rp” [OCR artifact].
handle with care

• Alas, introducing automatic spelling correction (fuzzy matching) to entity recognition often requires the introduction of white word lists to avoid problems.
  – herein → heroin
  – aspiring → aspirin
  – cranium → uranium
  – ability → abilify

• More aggressive correction leads to more problems:
  – “be that the line” → methantheline
  – “park on a zone” → parconazole
To quantify the benefits of automatic spelling correction to “real world” chemical text mining we analysed the first 28 weeks of US patent grants from 2012.

This corresponds to 145,473 documents, issued between 3rd January 2012 and 10th July 2012.

A total of 4,061,670 IUPAC-like systematic names were identified, with 1,816,317 unique patent/name pairs.

OPSIN interprets 3,722,399 and 1,647,402.
total molecule entities

All Extracted Molecule Entities

- Correct: 341,166,500, 84%
- Simple: 55,253,300, 14%
- D=1: 9,747,200, 2%

OPSIN Interpreted Molecule Entities

- Correct: 320,019,300, 86%
- Simple: 43,448,700, 12%
- D=1: 8,771,900, 2%

Using correction retrieves ~19% more entities and ~16% more OPSIN recognizable names.
The effect is more pronounced with patent-cmpnd pairs with a +27% improvement over no correction.
influence on n2s software (opsin)

<table>
<thead>
<tr>
<th>Interpretation</th>
<th>Count</th>
<th>Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Interpretable</td>
<td>120673</td>
<td>19.23%</td>
</tr>
<tr>
<td>Before not After</td>
<td>11814</td>
<td>1.88%</td>
</tr>
<tr>
<td>After not Before</td>
<td>249399</td>
<td>39.74%</td>
</tr>
<tr>
<td>Same Before/After</td>
<td>187926</td>
<td>29.95%</td>
</tr>
<tr>
<td>Different Before/After</td>
<td>57702</td>
<td>9.20%</td>
</tr>
<tr>
<td>Total</td>
<td>627514</td>
<td>100.00%</td>
</tr>
</tbody>
</table>

Although some valid names are lost by correction, overall the effect is overwhelmingly positive.
break-down of edit operations

<table>
<thead>
<tr>
<th>Edit Operation</th>
<th>Count</th>
<th>Fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deletion</td>
<td>390561</td>
<td>48.69%</td>
</tr>
<tr>
<td>Insertion</td>
<td>222716</td>
<td>27.77%</td>
</tr>
<tr>
<td>Substitution</td>
<td>186142</td>
<td>23.21%</td>
</tr>
<tr>
<td>Transposition</td>
<td>2671</td>
<td>0.33%</td>
</tr>
<tr>
<td>Total</td>
<td>802090</td>
<td>100.00%</td>
</tr>
</tbody>
</table>
Some improvement is seen with drug dictionaries, but there’s little benefit in fixing simple OCR issues.
target dictionary entries

• Simple correction of ChEMBL protein target names
  prostaglandin H-2 synthase-1 → prostaglandin H2 synthase 1
  Alanine amino-transferase → Alanine aminotransferase
  acetyl cholinesterase → acetylcholinesterase
  cyclooxygenase-2 → cyclooxygenase-2
  MAP kinase ERK-2 → MAP kinase ERK2
  HEC-GLCNAC-6-ST → HEC-GLCNAC6ST
  Herne Oxygenase → Heme Oxygenase
  Prealbumin → Prealbumin
  p110- delta → p110delta
non-word spelling correction

• Automatic correction technology can also be applied to entities other than words or IUPAC-like chemical nomenclature.
• For example, Chemical Abstract Service’s registry numbers.
Two to seven digits, followed by a hyphen, two digits, a hyphen and a final check digit
   - e.g. 7732-18-5

Regular Expression: \(((1-9)\d{2,5})\|([5-9]\d))-\d\d-\d\d
Cas check digit calculation

- More generally CaffeineFix’s finite state machines can do limited processing...
- The final check digit of a CAS number is calculated by series term summation modulo 10.
- The last digit time 1, the previous digit times 2, the previous digit times 3, and computing the sum modulo 10.
- The CAS number for water is 7732-18-5.
- The checksum 5 is calculated as \( (1 \times 8 + 2 \times 1 + 3 \times 2 + 4 \times 3 + 5 \times 7 + 6 \times 7) \mod 10 = 5 \).
Fsm for matching cas check digits
cas number correction example

• 7732-18-8? Did you mean...
  – 7732-18-5
  – 7732-11-8
  – 77328-18-8
  – 7733-18-8
  – 77342-18-8
  – 77392-18-8
  – 71732-18-8
  – 76732-18-8
  – 97732-18-8
take home message

• Adding advanced automatic chemical spelling correction to an annotation pipeline typically improves recall by about 20-40%.

  – Sorel Muresan, “Automated Spelling Correction to Improve Recall Rates of Name-to-Structure Tools for Chemical Text Mining”, 2011 ChemAxon EUGM.
acknowledgements

• Daniel Lowe, NextMove Software.
• Sorel Muresan and Paul Hongxing Xie, AstraZeneca.

• Thank you for your time.
• Any questions?