Advances in Automatic Chemical Spelling Correction

Roger Sayle and Daniel Lowe

NextMove Software

Cambridge, UK

ACS National Meeting, Philadelphia, USA 19th August 2012

EXAMPLE SPELLING ERRORS

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

Incorrect Name	US Patent No.	Issue Date
pryimidine	8093264	10th January 2012
pyrmidine	8097728	17th January 2012
pyrimdine	8114996	14th February 2012
pyrimidne	8129897	6th March 2012
pyrimidinc	8148339	3rd April 2012
pyridmidine	8158627	8th May 2012

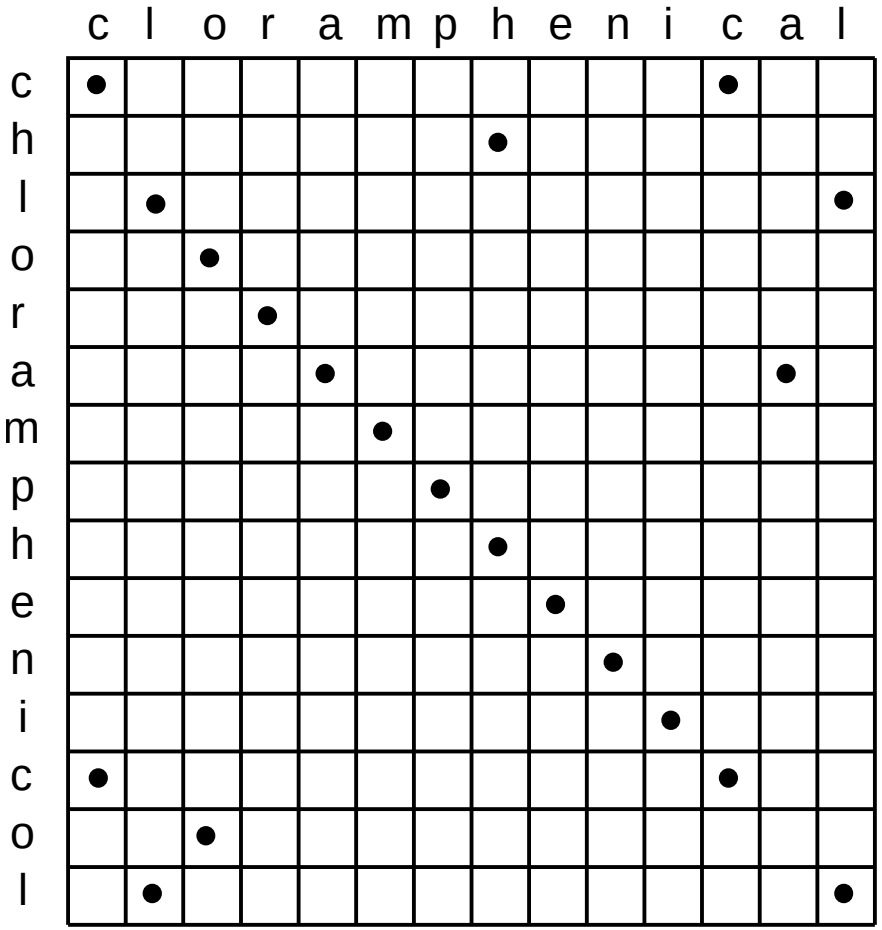
PREVIOUS WORK

- Roger Sayle, Paul Hongxing Xie and Sorel Muresan, “Improved Chemical Text Mining of Patents with Infinite Dictionaries and Automatic Spelling Correction”, Journal of Chemical Information and Modeling, Vol. 52, No. 1, pp. 51-62, 2012.
- G.H. Kirby, M.R. Lord, J.D. Rayner, “Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 6. (Semi)automatic name correction, Journal of Chemical Information and Computer Science, Vol. 31, pp. 153-160, 1991.

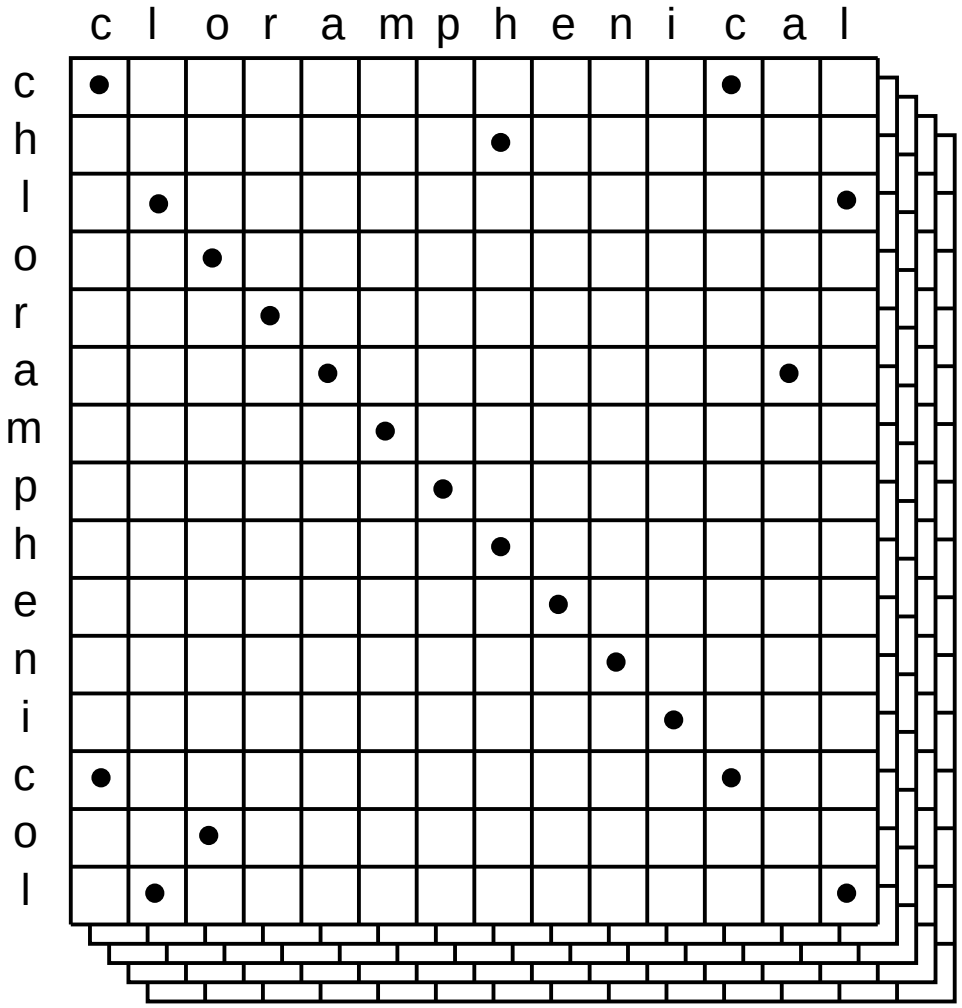
STRING EDIT DISTANCE

- 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



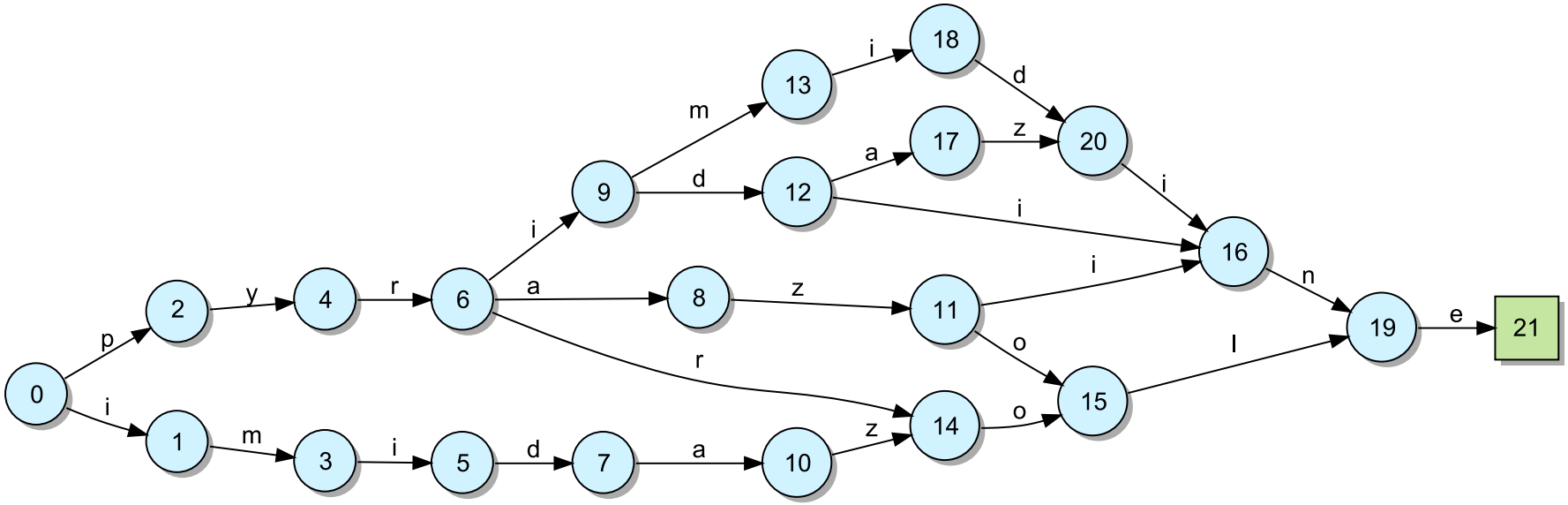
NEEDLEMAN-WUNSCH-SELLERS



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.

DICTIONARIES AS AUTOMATA



- 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.

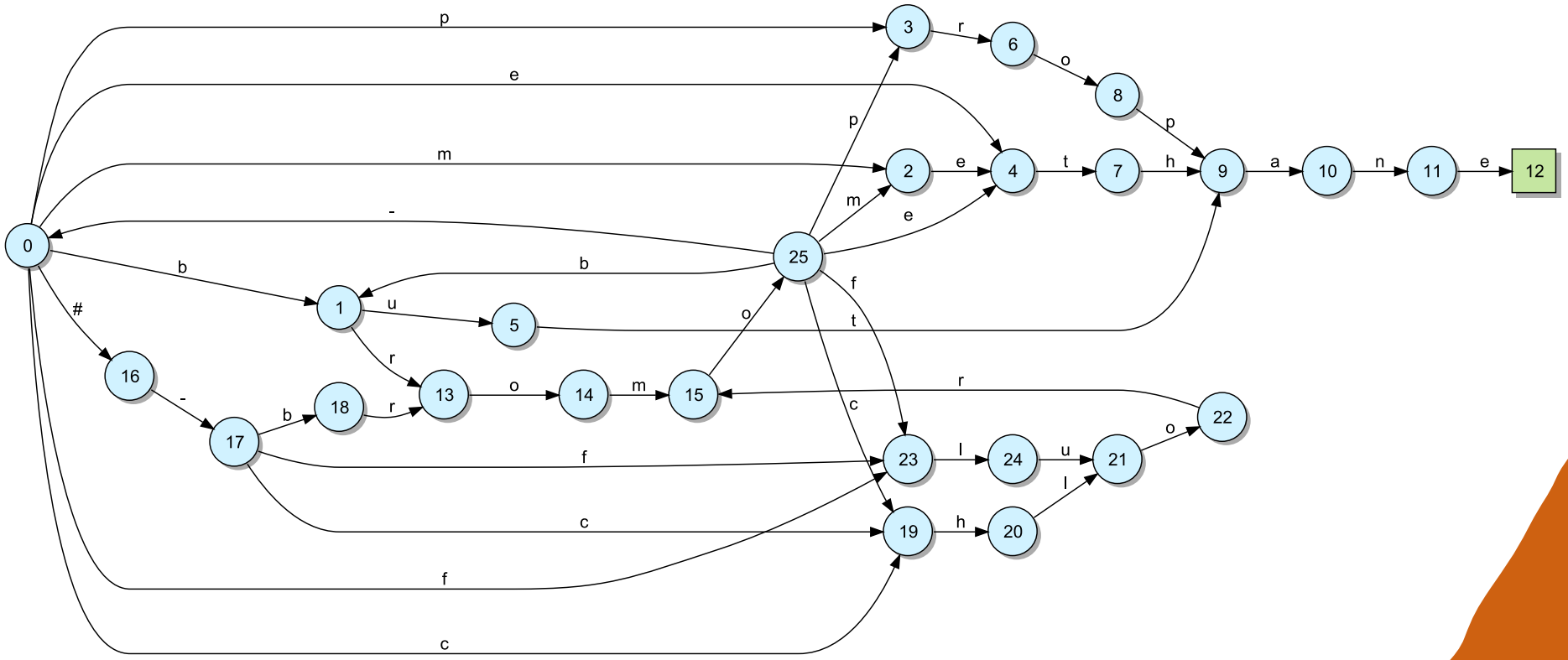
```
alk := "meth" | "eth" | "prop" | "but"
parent := alk "ane"
subst := "bromo" | "chloro" | "fluoro"

locant := "1" | "2" | "3" | "4" /* any digit */
prefix := [ prefix "-" ] [ loc "-" ] subst
        | [ prefix ] subst
name := [ prefix [ "-" ] ] parent
```

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-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...
  - 99.15% (232144/234142) names in the NCI00 database.
  - 95.28% (67995/71367) names in the Maybridge catalogue.
  - 95.27% (25890/48167) names in the Keyorganics catalogue.
- These figures are comparable to name-to-structure conversion rates on these names.

SPELLING CORRECTION EXAMPLES

- lH-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 “I” (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 “φ” by “rp” [OCR artifact].

HANDLE WITH CARE

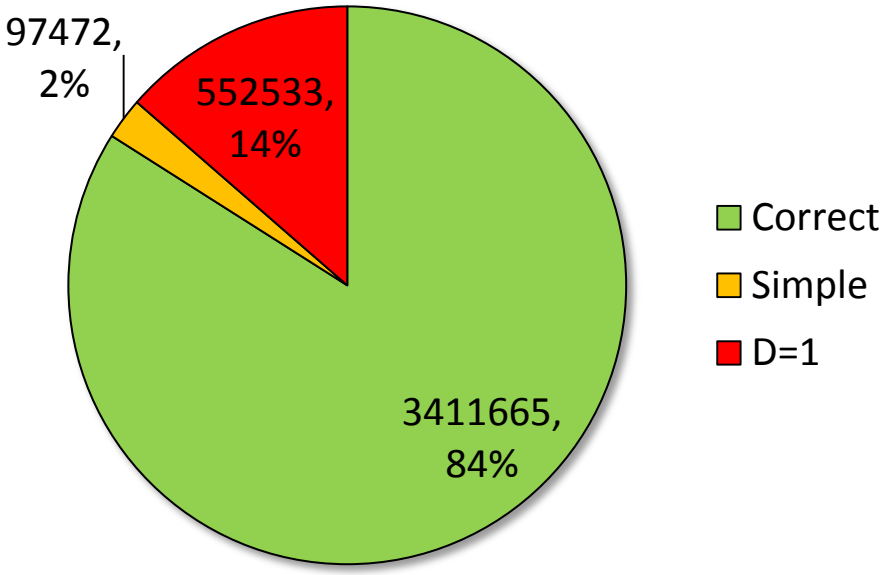
- Alas, introducing automatic spelling correction (fuzzy matching) to entity recognition often requires the introduction of white word lists to avoid problems.
  - herein → heroin
  - aspiring → aspirin
  - cranium → uranium
  - ability → abilify
- More aggressive correction leads to more problems:
  - “be that the line” → methantheline
  - “park on a zone” → parconazole

BENCHMARKING AND ANALYSIS

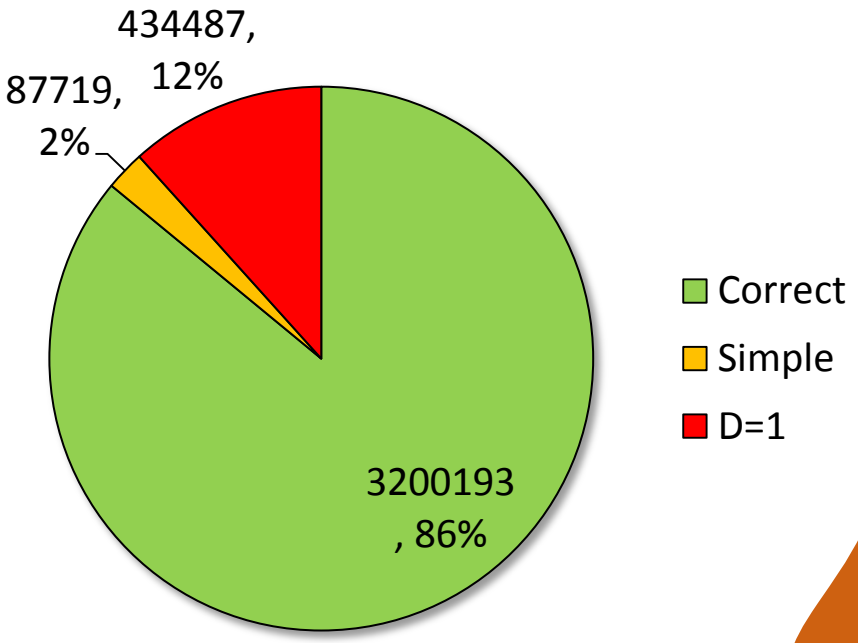
- To quantify the benefits of automatic spelling correction to “real world” chemical text mining we analysed the first 28 weeks of US patent grants from 2012.
- This corresponds to 145,473 documents, issued between 3rd January 2012 and 10th July 2012.
- A total of 4,061,670 IUPAC-like systematic names 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



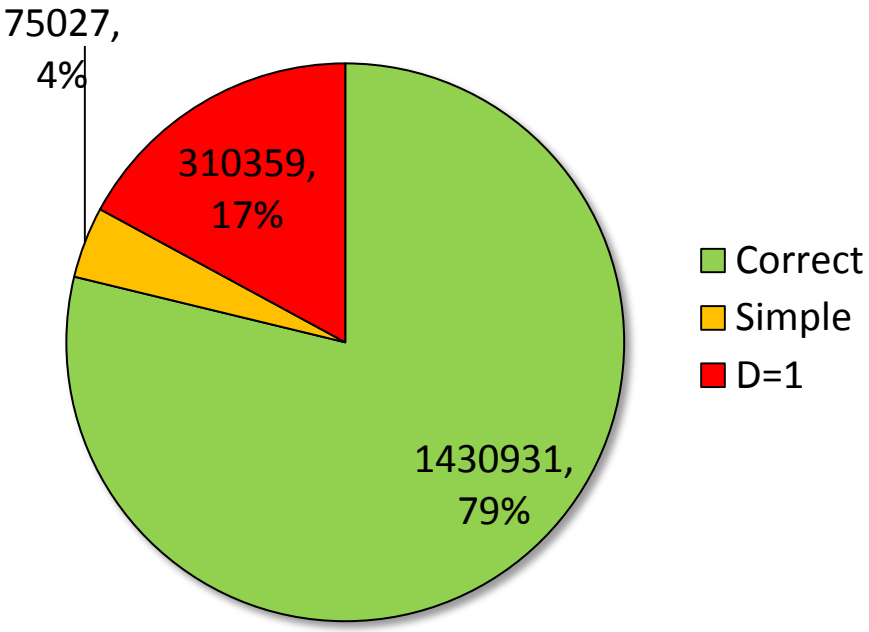
OPSIN Interpeted Molecule Entities



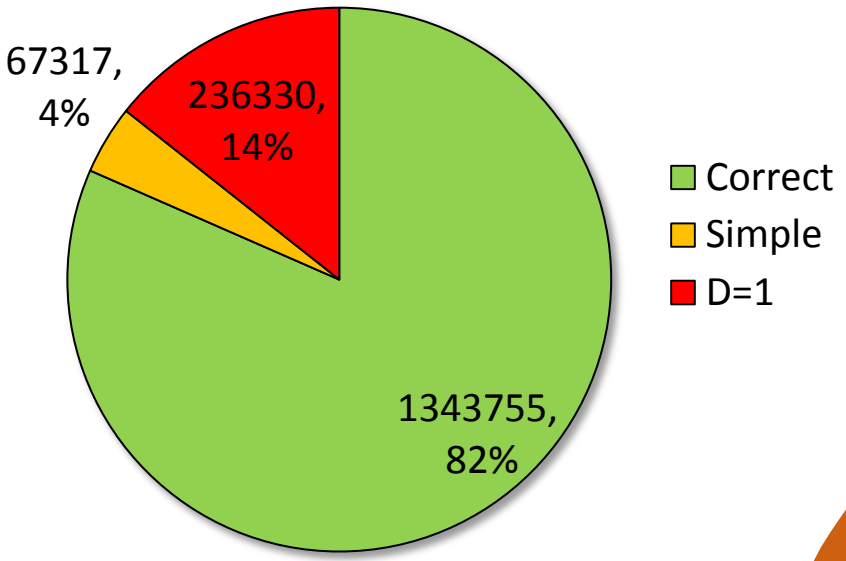
Using correction retrieves ~19% more entities and ~16% more OPSIN recognizable names.

UNIQUE MOLECULE ENTRIES

All Extracted Molecule Entities



OPSIN Interpeted Molecule Entities



The effect is more pronounced with patent-cmpnd pairs with a +27% improvement over no correction.

INFLUENCE ON N2S SOFTWARE (OPSIN)

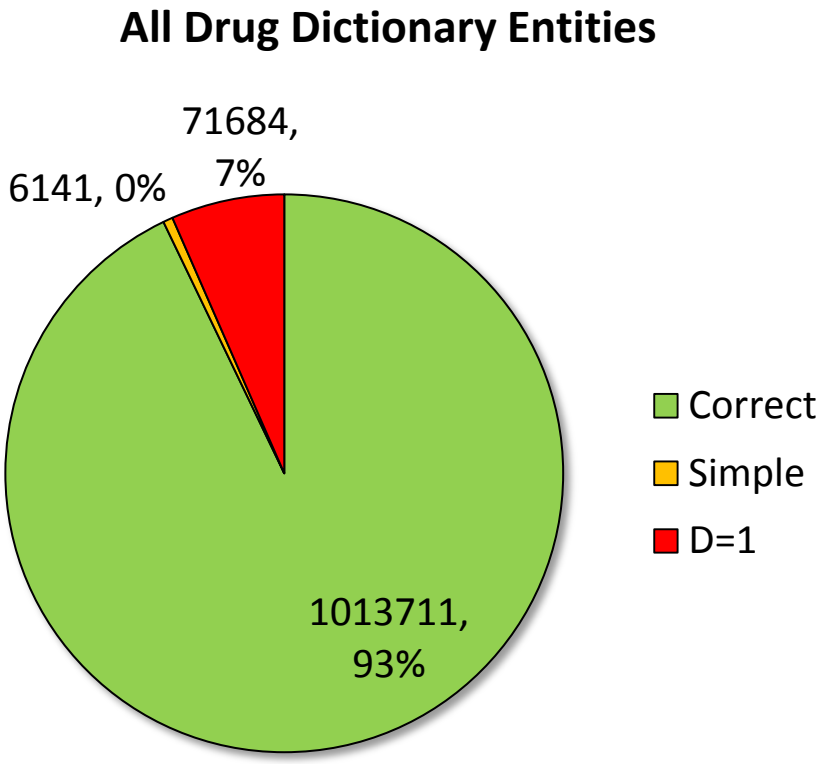
Interpretation	Count	Fraction
Not Interpretable	116216	17.88%
Before not After	11779	1.81%
After not Before	270453	41.61%
Same Before/After	181693	27.95%
Different Before/After	69864	10.75%
Total	650005	100.00%

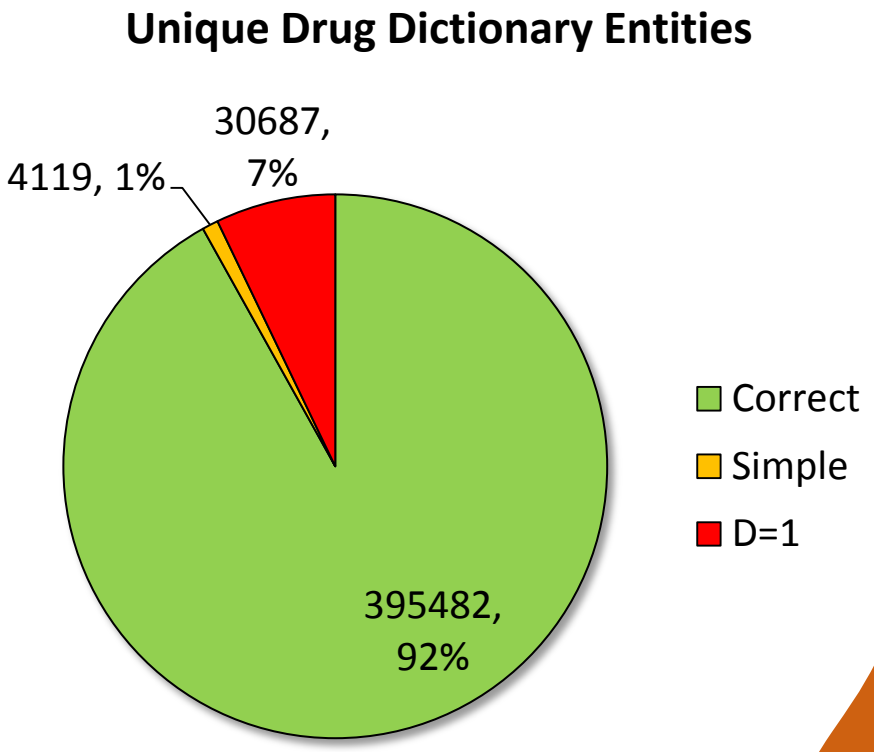
Although some valid names are lost by correction, overall the effect is overwhelmingly positive.

BREAK-DOWN OF EDIT OPERATIONS

Edit Operation	Count	Fraction
Deletion	392324	47.50%
Insertion	232493	28.15%
Substitution	198438	24.03%
Transposition	2670	0.32%
Total	825,925	100.00%

DRUG DICTIONARY ENTITIES





Some improvement is seen with drug dictionaries, but there's little benefit in fixing simple OCR issues.

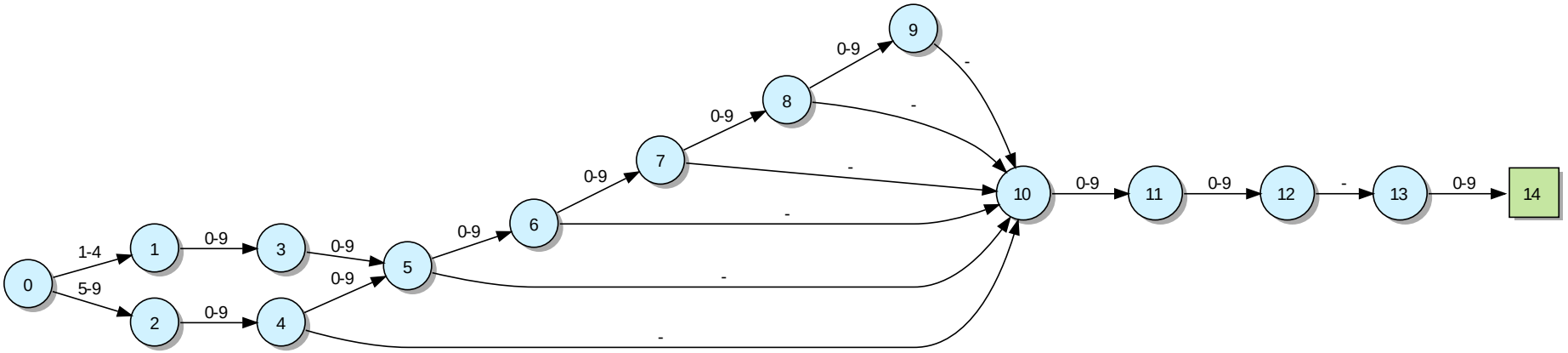
TARGET DICTIONARY ENTRIES

- Simple correction of ChEMBL protein target names
- prostaglandin H- 2 synthase- 1 → prostaglandin H2 synthase 1
- Alanine amino-transferase → Alanine aminotransferase
- acetyl cholinesterase → acetylcholinesterase
- cyclooxy-genase-2 → cyclooxygenase-2
- MAP kinase ERK-2 → MAP kinase ERK2
- HEC-GLCNAC-6-ST → HEC-GLCNAC6ST
- Herne Oxygenase → Heme Oxygenase
- Prealburnin → 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.

CAS REGISTRY NUMBER GRAMMAR

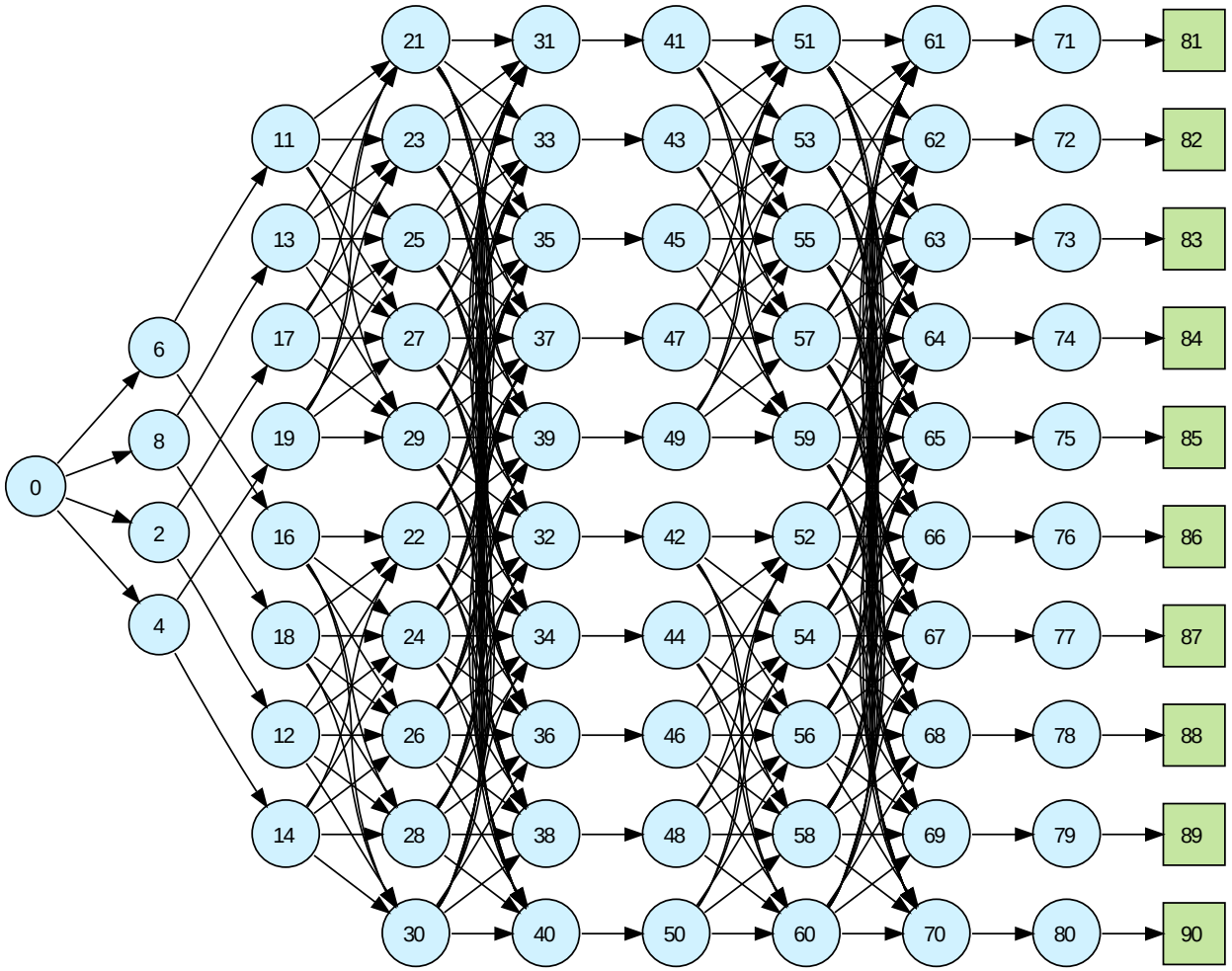


- Two to seven digits, followed by a hyphen, two digits, a hyphen and a final check digit
  - e.g. 7732-18-5
- Regular Expression: (([1-9]\d{2,5})|([5-9]\d))-\d\d-\d

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 (1x8 + 2x1 + 3x2 + 4x3 + 5x7 + 6x7) 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%.
  - Andrew Hinton, “Benchmarking ChemAxon’s Name-to-Structure batch tool on Patent Data”, 2011 ChemAxon EUGM, Budapest.
  - 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?