



IMPROVED CHEMICAL TEXT MINING OF PATENTS USING AUTOMATIC SPELLING CORRECTION AND INFINITE DICTIONARIES

Roger Sayle¹, Paul-Hongxing Xie², Plamen Petrov², Jon Winter³ and Sorel Muresan²

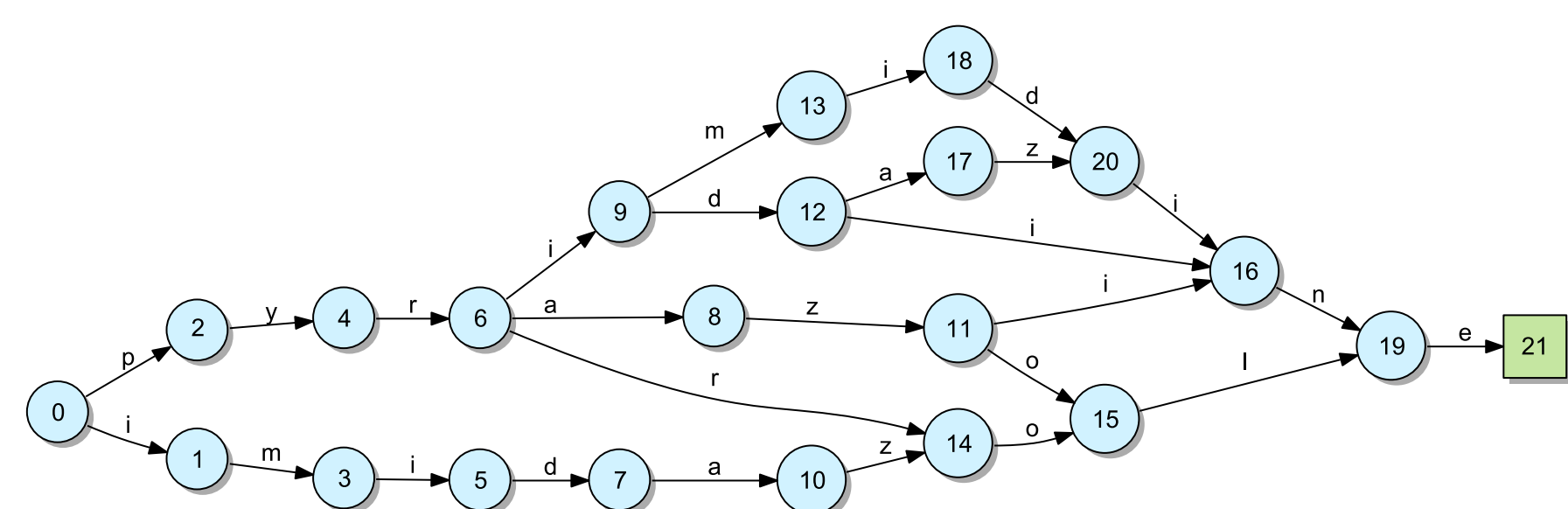
¹ NextMove Software Ltd, Cambridge, UK ² AstraZeneca, Mölndal, Sweden ³ AstraZeneca, Alderley Park, UK

1. Abstract

The text mining of patents for chemical structures of pharmaceutical interest poses a number of unique challenges not encountered in other fields of text mining. Unlike fields such as bioinformatics where the number of terms of interest is enumerable and static, systematic chemical nomenclature can describe an infinite number of molecules. Hence the dictionary techniques that are commonly used for gene names, diseases, species etc. have limited utility when searching for novel therapeutic compounds in patents. Additionally, the length and composition of IUPAC-like names makes them more susceptible to typographical problems; OCR failures, human errors and hyphenation and line breaking issues. This work describes a novel technique, called CaffeineFix, designed to efficiently and correctly identify chemical names in free text, even in the presence of typographical errors. This forms a pre-processing pass, independent of the name-to-structure software used, and is shown to greatly improve results in our study.

2. Efficient Dictionaries

The CaffeineFix algorithm efficiently represents the set or dictionary of entities to recognize as a minimal finite state machine (FSM). This encoding allows it to very efficiently match a string against very large dictionaries; significantly more efficient than the hashing or search tree based methods used by Java or C++'s STL library. The FSM below demonstrates the encoding a dictionary for small nitrogen containing heterocycles, containing pyrrole, pyrazole, imidazole, pyridine, pyridazine, pyrimidine and pyrazine.

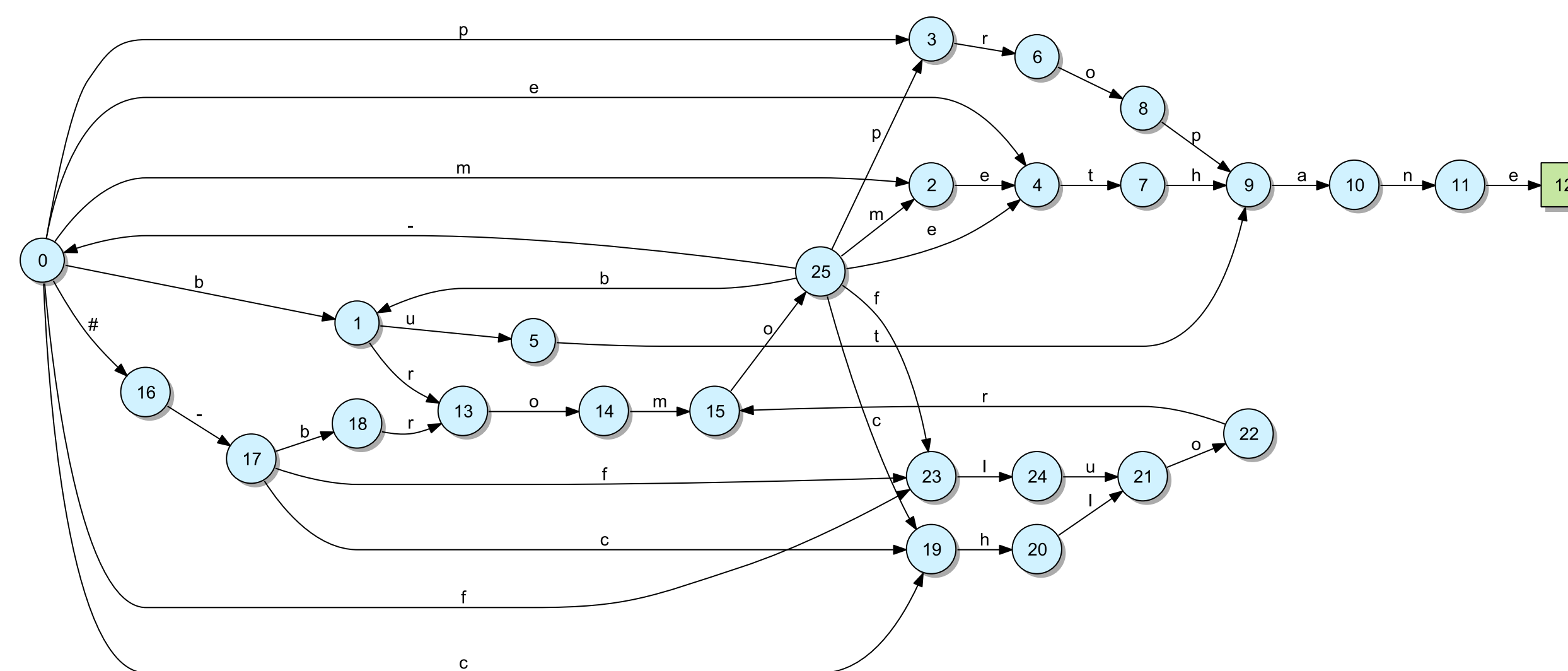


3. Improved Tokenization

A major benefit of using FSMs for chemical text mining is improved tokenization (the splitting of free text into words/terms). Chemical names can contain spaces, hyphens, digits, commas, parentheses, brackets, braces, apostrophes, periods, superscripts and Greek characters that confound most NLP tools. This is made harder still by typos, OCR errors, line breaks and hyphenation, XML/HTML tags, line and page numbers. The ability to FSMs to efficiently recognize valid prefixes, allows characters such as spaces and commas to be delimiters in some contexts but part of the recognized entity in others.

4. Infinite Dictionaries (Grammars)

A novel advance over previous spelling checking implementations is to observe that FSM representations can encode an infinite number of terms by permitting backward edges, and using stacks (push down automata) for balancing brackets in context-free (IUPAC-like) grammars. Consider a simple chemical grammar consisting of optionally halo substituted alkanes; where the prefixes "bromo", "chloro" and "fluoro" may be repeatedly applied to "methane", "ethane", "propane" or "butane".



This FSM can recognize an infinite number of strings include "methane", "chloroethane", "2-bromo-propane", "chloro-bromo-methane" and so on.

5. Systematic IUPAC/CAS-like Grammar

To recognize the majority of nomenclature used in medicinal chemistry, CaffeineFix's current IUPAC dictionary (FSM) contains 452,126 states. This covers 98.97% of the OpenEye Lexichem generated names for the NCI00 database, and 91.23% of the 71K names in the Maybridge 2003 catalogue.

6. Automatic Spelling Correction

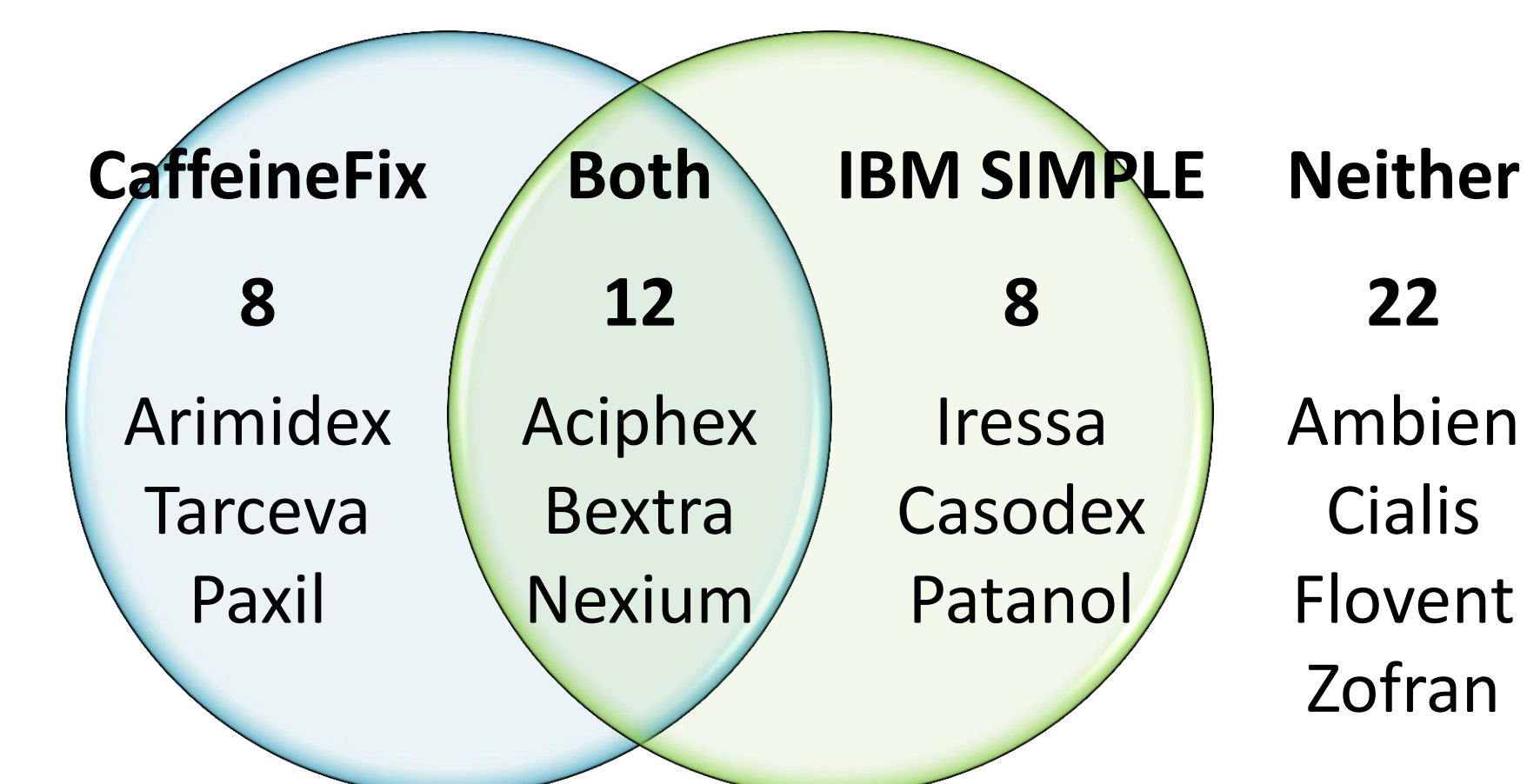
Another major benefit of using FSMs in CaffeineFix is the ability to perform "fuzzy" matching, returning all possible terms within a limited *Levenshtein* (or string-edit) distance of a query string. By backtracking over the FSM, it is possible to determine the minimum number of single character insertions, deletions or substitutions required to transform one string into another. When only a single unique "correction" can be found within a given radius, the misspelling may be corrected automatically. CaffeineFix suggests "1,2-dichlorobenzene" to replace "12-dichlorobenzene", "dodec-2-ene" for "didec-2-ene" and "spiro[2.3]hexane" for "spiro[2.2]hexane". Additionally penalties/distances may be parameterized (as in bioinformatics) such that homoglyphic substitutions found in OCR (between "1", "l" and "I", between "0" and "O" or insertion of "
") cost less than other changes.

7. Recall Benchmark Results

The methods were applied to mining IBM's text database of 12 million US, European and world patents. Non-English abstracts were translated using OpenEye's Lexichem translation functionality. The analysis yielded a total of 13,523,284 unique chemical names, including 7,262,798 systematic names. Using a suite of name-to-structure converters including Lexichem, ChemAxon and OPSIN allowed 92.2% of these names to be converted to structures. In total, 5,805,172 unique canonical SMILES were indexed.

8. Precision Benchmark Results

To assess the quality of structures extracted, a set of 50 US patents denoting top selling drugs was used as a benchmark. These included US4255431 (Losec), US4681893 (Lipitor), US4847265 (Plavix), US6566360 (Levitra) and so on. The objective of the benchmark was to report the highest Tanimoto similarity, using MACCS 166-bit keys, between the query "drug" and the compounds extracted from its patent.



CaffeineFix (D=1) and Lexichem equalled IBM's SIMPLE annotator with 20/50 exact matches, and ahead of OSCAR/OPSIN's 17. Multiple name-to-structure programs and more aggressive spelling correction produced better results.

9. Conclusions

CaffeineFix's use of finite state machines allows it to efficiently recognize an infinite number systematic IUPAC-like chemical names in free text, even in the presence of OCR and other typographical errors. This is shown to improve the extraction of relevant structures from pharmaceutical patents.

10. Bibliography

1. Roger Sayle, "Foreign Language Translation of Chemical Nomenclature by Computer", *JCIM*, Vol. 49, No. 3, pp. 519-530, 2009.
2. James Rhodes, Stephen Boyer, Jeffrey Kreule, Ying Chen and Patricia Ordonez, "Mining Patents using Molecular Similarity Search", *Pacific Symposium on Biocomputing*, Vol. 12, pp. 304-315, 2007.
3. Ithipol Suriyawongkul, Chris Southan and Sorel Muresan, "The Cinderella of Biological Data Integration: Addressing the Challenges of Entity and Relationship Mining from Patent Sources", In *Data Integration in the Life Sciences*, Springer 2010.