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

ROGER SAYLE, PAUL-HONGXING XIE, PLAMEN PETROV, JON WINTER AND SOREL MURESAN

NEXTMOVE SOFTWARE, SANTA FE, NEW MEXICO, USA
& ASTRAZENECA R&D, MOLNDAL, SWEDEN

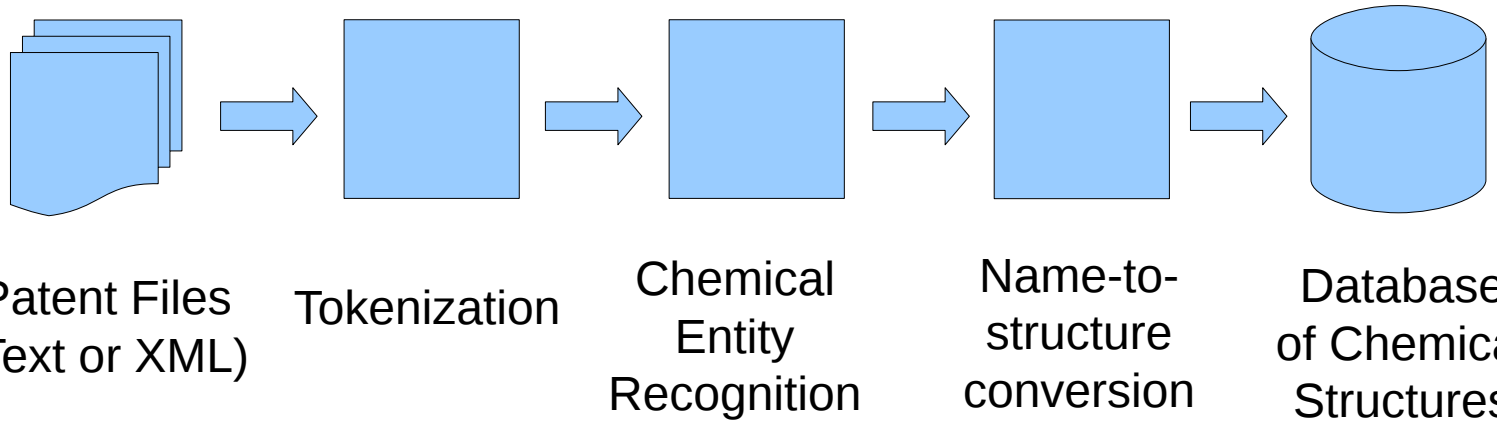
OVERVIEW

- Why?
  - Chemical Named Entity Extraction
- How?
  - Dictionary Matching Algorithms
  - Grammars and Infinite Dictionaries
  - Automatic Spelling Correction
  - Languages and Translation
- What?
  - Results and Conclusions

MOTIVATION

- “The biggest cause of missing compounds when extracting chemical entities from text is the presence of typographical errors: human errors, OCR failures, hyphenation and multiple line issues, etc.”

TRADITIONAL TEXT MINING PIPELINE



- Typically, text mining is organized as a pipeline with the text initially being broken into segments, perceived as chemicals and finally converted to SMILES/SD files.

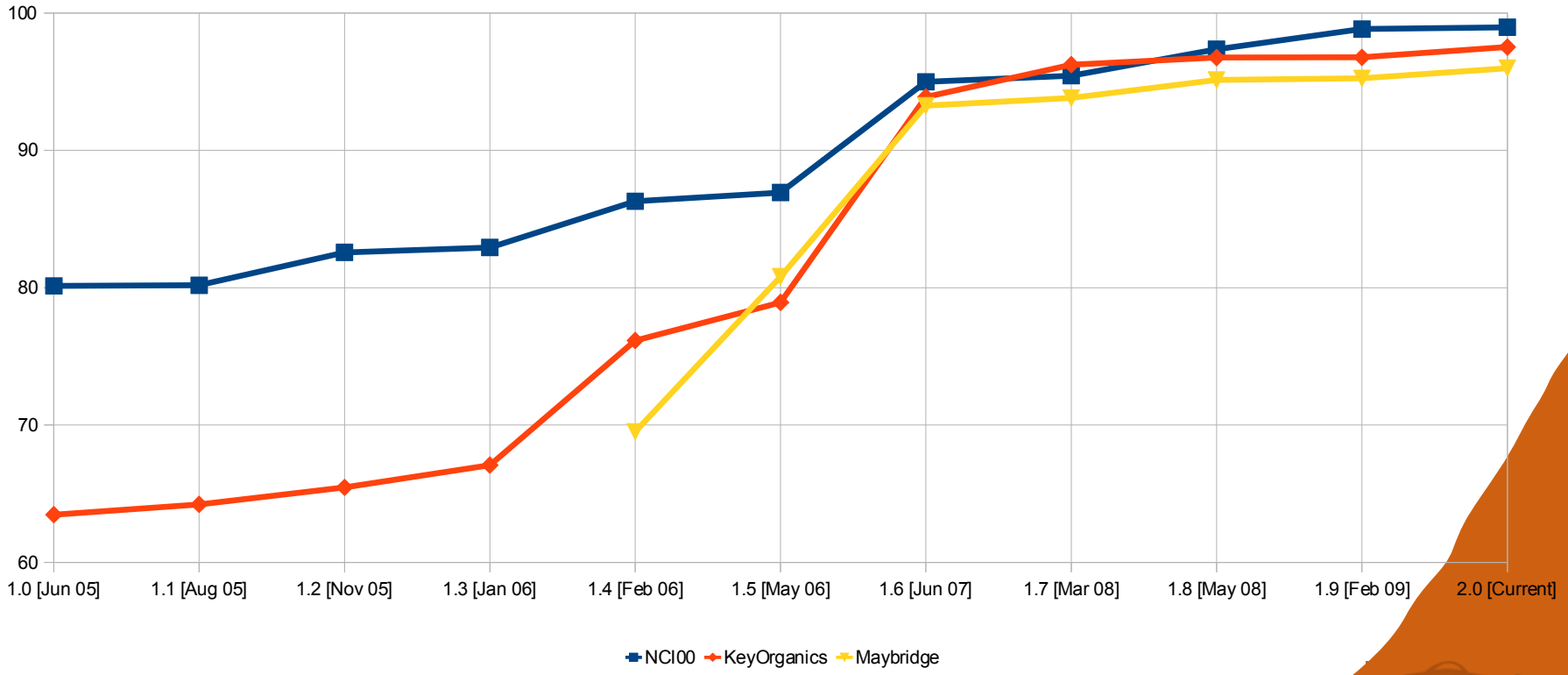
CHEMICAL NAME SEGMENTATION

- Determining the start and end of IUPAC-like names in free text is a tricky business.
- Chemical names can contain whitespace, hyphens, commas, parenthesis, brackets, braces, apostrophes, superscripts, greek characters, digits and periods.
- This is made harder still by typos, OCR errors, hyphenation, linefeeds, XML tags, line and page numbers and similar noise.

EVOLUTION OF OPENEYE'S LEXICHEM

Version	NCI00	Keyorganics	Maybridge
2.0 [current]	98.96%	97.52%	95.98%
1.9 [Feb 09]	98.83%	96.77%	95.24%
1.8 [May 08]	97.36%	96.75%	95.12%
1.7 [Mar 08]	95.44%	96.24%	93.81%
1.6 [Jun 07]	94.99%	93.89%	93.25%
1.5 [May 06]	86.29%	78.93%	80.80%
1.4 [Feb 06]	86.93%	76.15%	69.51%
1.3 [Jan 06]	82.93%	67.08%	
1.2 [Nov 05]	82.57%	65.46%	
1.1 [Aug 05]	80.18%	64.22%	
1.0 [Jun 05]	80.13%	63.47%	

EVOLUTION OF OPENEYE'S LEXICHEM



DIFFERENCE IN THEORY VS. PRACTICE

- Unfortunately, a major usage of Name-to-structure sofware, such as Lexichem, is in text mining and chemical named entity extraction from patents and on-line web documents.
- In this use-case, the performance is not limited by the chemistries supported by name-to-structure conversion, but the high rate of typos and lexicographic errors.

THE MYTH OF DICTIONARIES

- Unfortunately, the dictionary-based approaches commonly used in other fields of text mining are of little use in the analysis of chemical patents.
- InfoChem's product literature describes their “huge dictionary containing more than 30 million unique entries” that underpins both their $IC_{ANNOTATOR}$ and IC_{N2S} products.

UNIQUENESS IN CHEMICAL PATENTS

- Google's PageRank and similar text clustering schemes rely on co-occurance of terms to identify similar documents.
- In pharmaceutical patents, it is the names that are unique in a document that are most relevant.
- Commonly occuring terms such as “water”, “acetone”, and so on, tend to be reagents of little importance or interest.

FUTILITY OF DICTIONARIES

- Whilst dictionaries are useful, especially for abbreviations, registry numbers, tradenames and other identifiers, they have limited utility in patent analysis.
- Chemical dictionaries are also difficult to curate and maintain; integrating data sources, eliminating duplicates, structure normalization and correcting errors.

QUALITY NOT QUANTITY

- When evaluating chemical text mining, the quality of which names are extracted are more important than the number of names.
- Poor segmentation and entity extraction can result in a number of small molecular fragments instead of a full molecule name.
- Judging chemical NER by molecule and/or fragment counts might not be optimal.

CATEGORIES OF CHEMICAL NAMES

M	Molecule	benzoic acid
D	Dictionary	ranitidine
R	Registry #	GW-409544
C	CAS Number	7732-18-5
E	Element	gold
P	Fragment	phenyl
A	Atom Fragment	chloro
Y	Polymer	polystyrene
G	Generic	alkane
N	Noise	formal

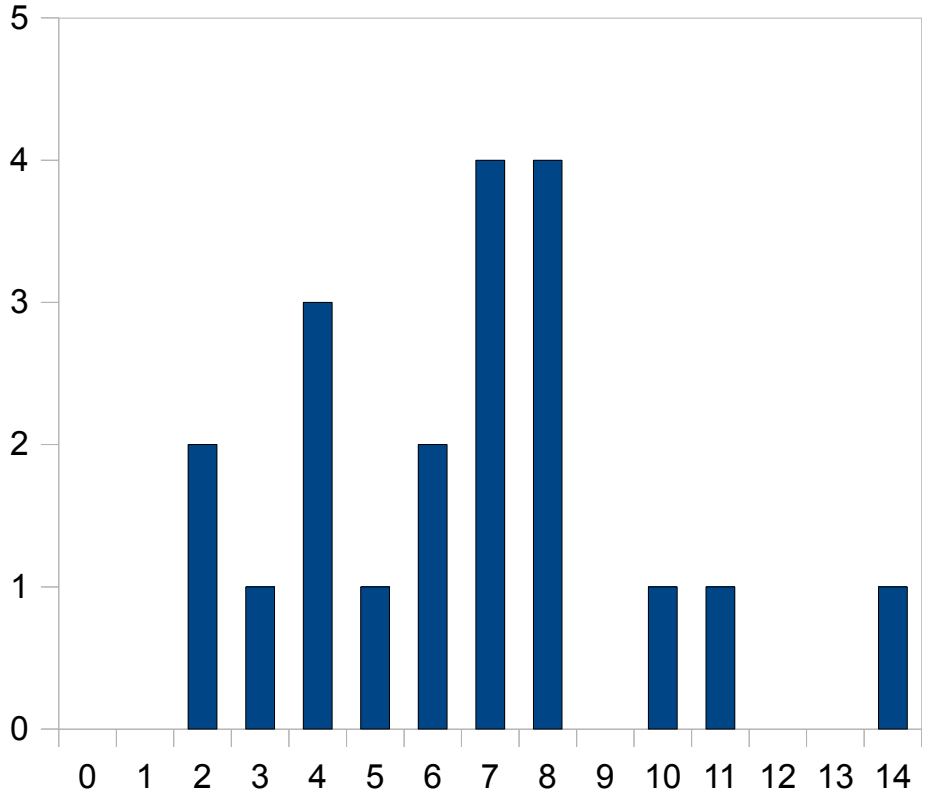
CASE STUDY: WO 2009126624

- By manual inspection, World Patent Office patent #2009126624 contains 20 names of pharmaceutical interest (“strategic”).
- None of these names are abstracted or indexed by IBM or SureChem/Nature.
- The primary cause is the large number of typos and OCR errors in the raw plain text.
- Lexichem gets all 20 “corrected” names.

LEVENSHTEIN DISTANCE (A.K.A. "STRING-EDIT DISTANCE")

- The minimum number of single character insertions, deletions and substitutions required to transform one string into another.
- Used as the mathematical basis of almost all sequence alignment and similarity algorithms in bioinformatics.
- Traditionally solved efficiently by dynamic programming algorithms.

PATENT OCR ERROR ANALYSIS



- Histogram of string edit distance (as a measure of error rate) of the 20 names described in WO2009126624.

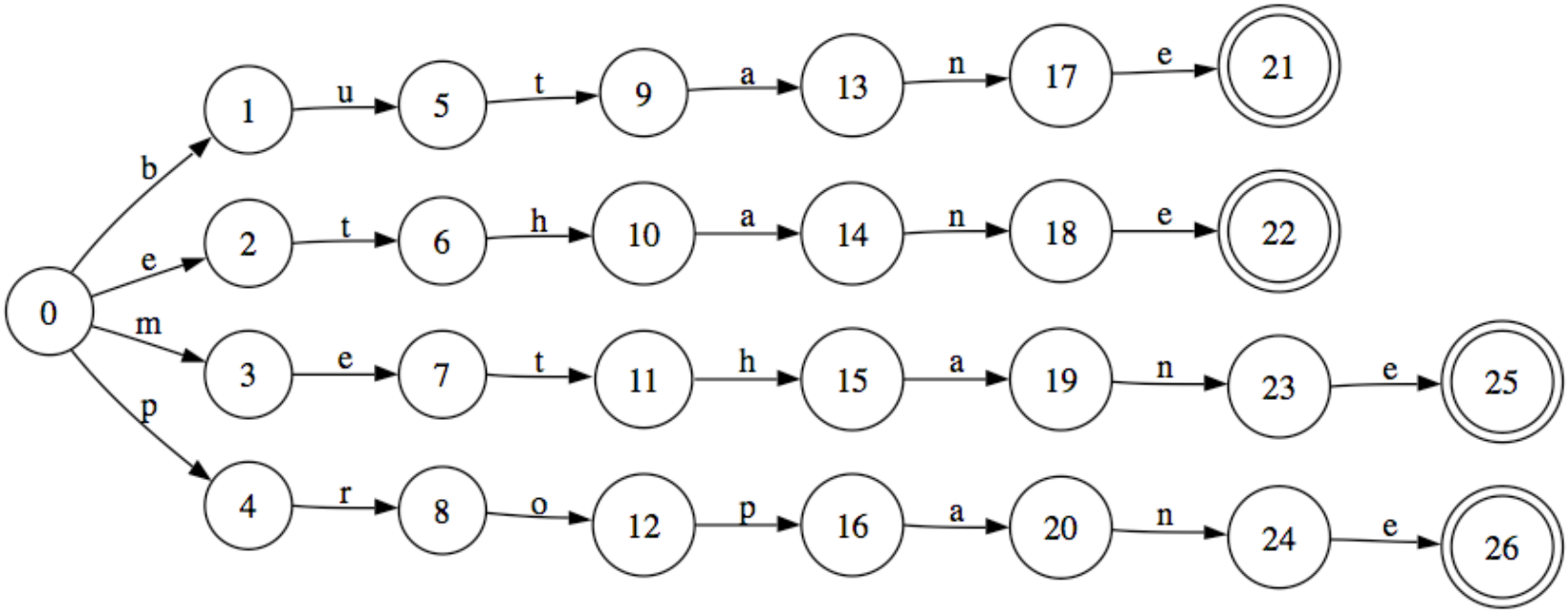
CAFFEINEFIX

- NextMove Software's CaffeineFix is intended to fill a niche opportunity as a chemical nomenclature aware automatic spell checker.
- As a pre-processing step in a pipeline, it can significantly improve the recall rates of name to structure tools in text mining applications; not just Lexichem, but also ACD/Name, ChemAxon, CambridgeSoft nam=struct, OPSIN, etc…

EXAMPLE CHEMICAL LEXICON

- Lower alkanes
  - Methane
  - Ethane
  - Propane
  - Butane

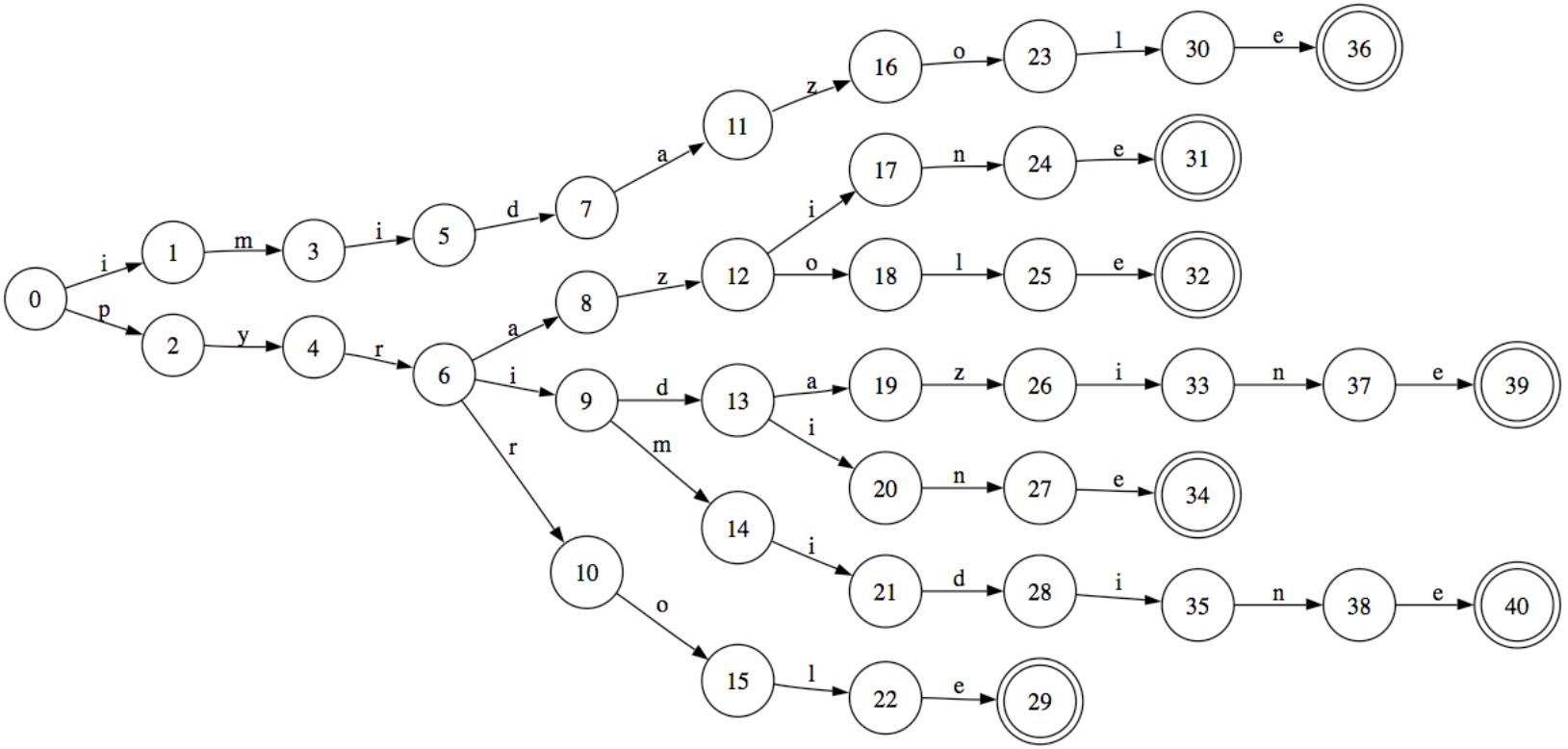
REPRESENTING LEXICONS AS TRIES



EXAMPLE CHEMICAL LEXICON

- Nitrogen containing heterocycles
  - Pyrrole
  - Pyrazole
  - Imidazole
  - Pyridine
  - Pyridazine
  - Pyrimidine
  - Pyrazine

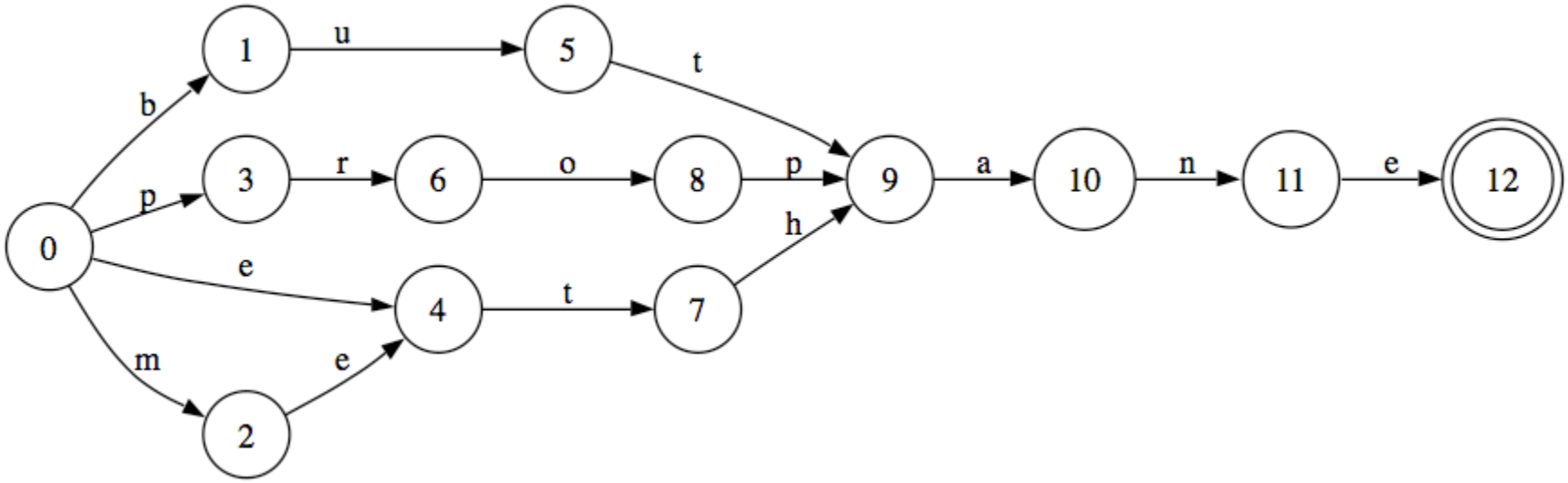
REPRESENTING LEXICONS AS TRIES



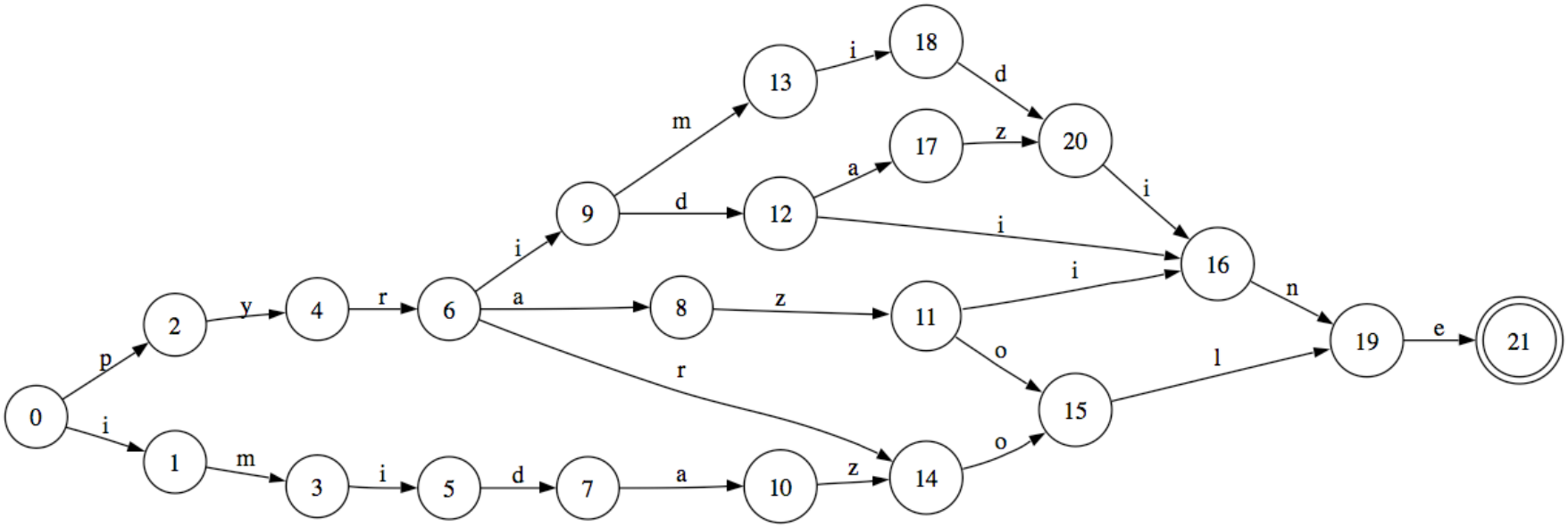
IS_MATCH IMPLEMENTATION

```
bool is_match(const unsigned char *ptr) {
  unsigned int state = 0;

  for(;;) {
    if (fsm[state].ch == *ptr) {
      ptr++;
      if (*ptr == '\0')
        return fsm[state].state;
      state = fsm[state].down;
    } else state = fsm[state].across;
    if (state == 0) return false;
  }
}
```

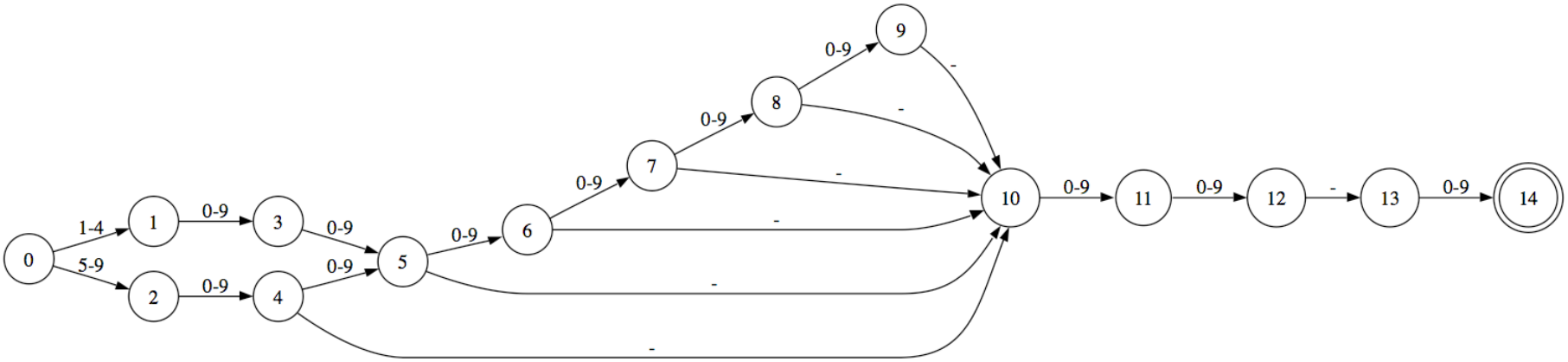
REPRESENTING LEXICONS AS DAGS



REPRESENTING LEXICONS AS DAGS



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
- RegExp: (([1-9]\d{2,5})|([5-9]\d))-\d\d-\d

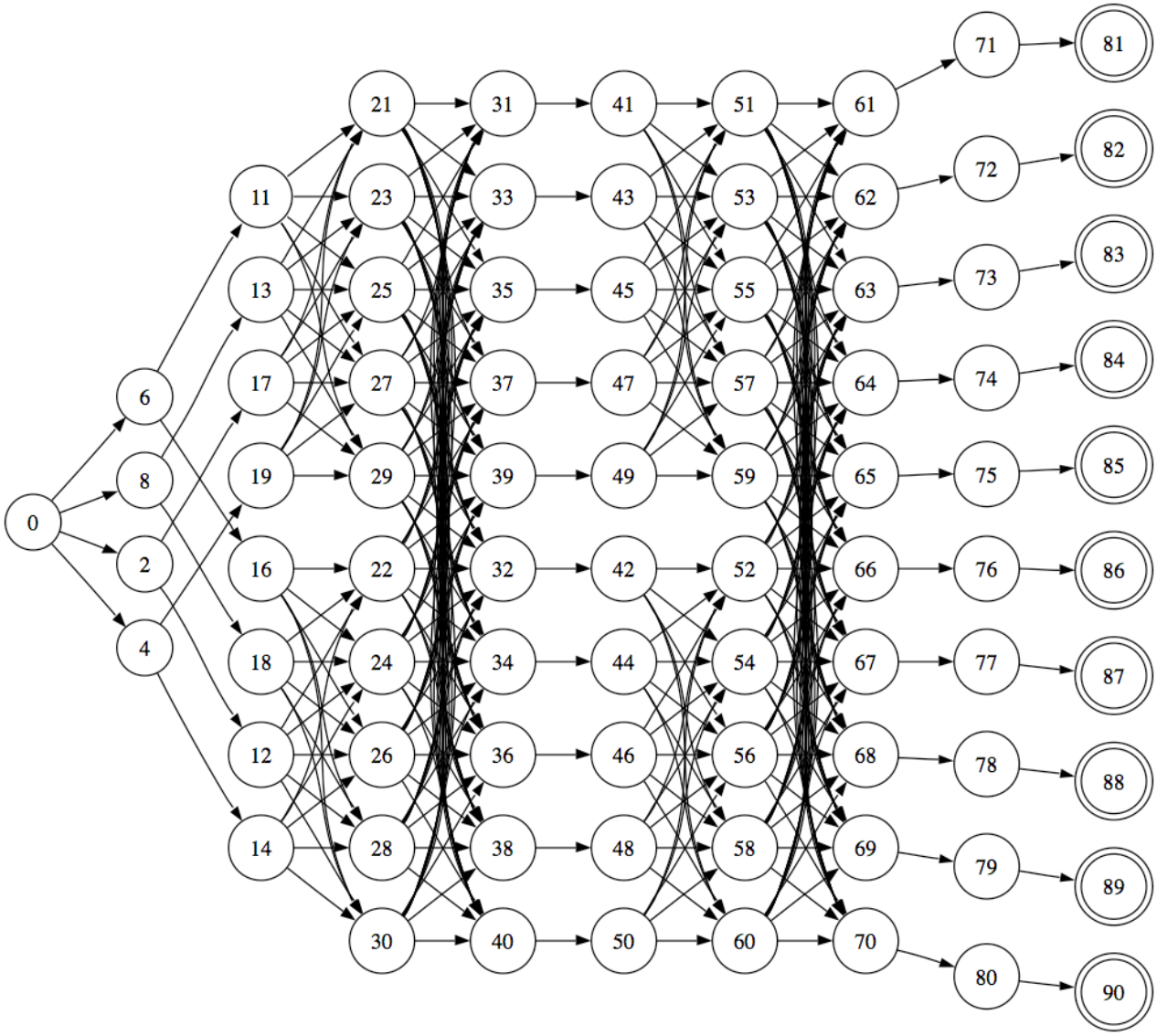
PHARMACEUTICAL REGISTRY NUMBERS

- Prefix: “A” | “AZ” | “BMY” | “GSK” | “LY” | …
- Number: \d{3-7}
- Suffix: (“.” \d) | [“a” .. “z”]
- Grammar: Prefix [“ ” | “-”] Number [Suffix]

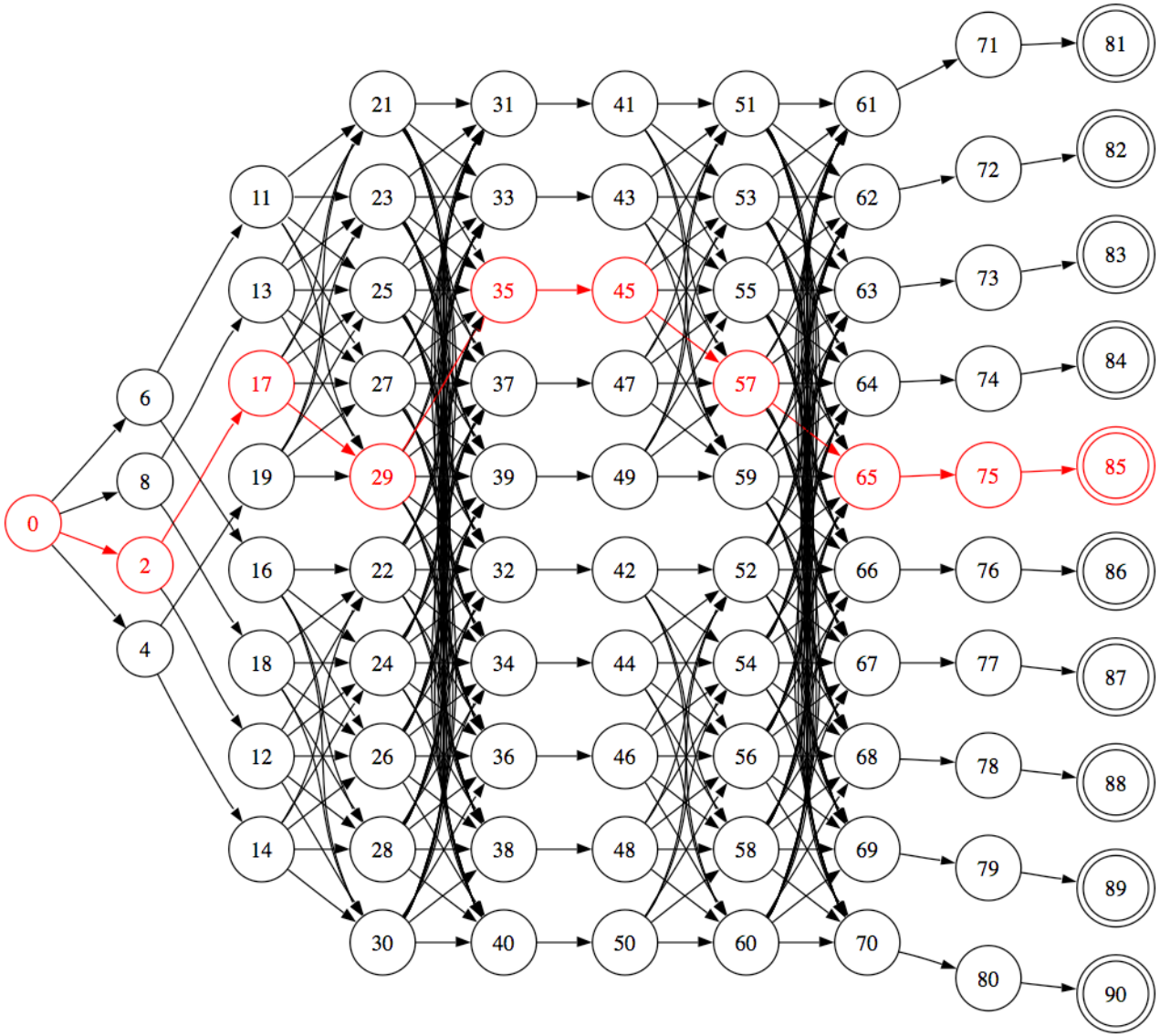
CAS CHECK DIGIT CALCULATION

- The final check digit is calcuated 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



FSM FOR MATCHING CAS CHECK DIGITS



EXAMPLE IUPAC-LIKE GRAMMAR

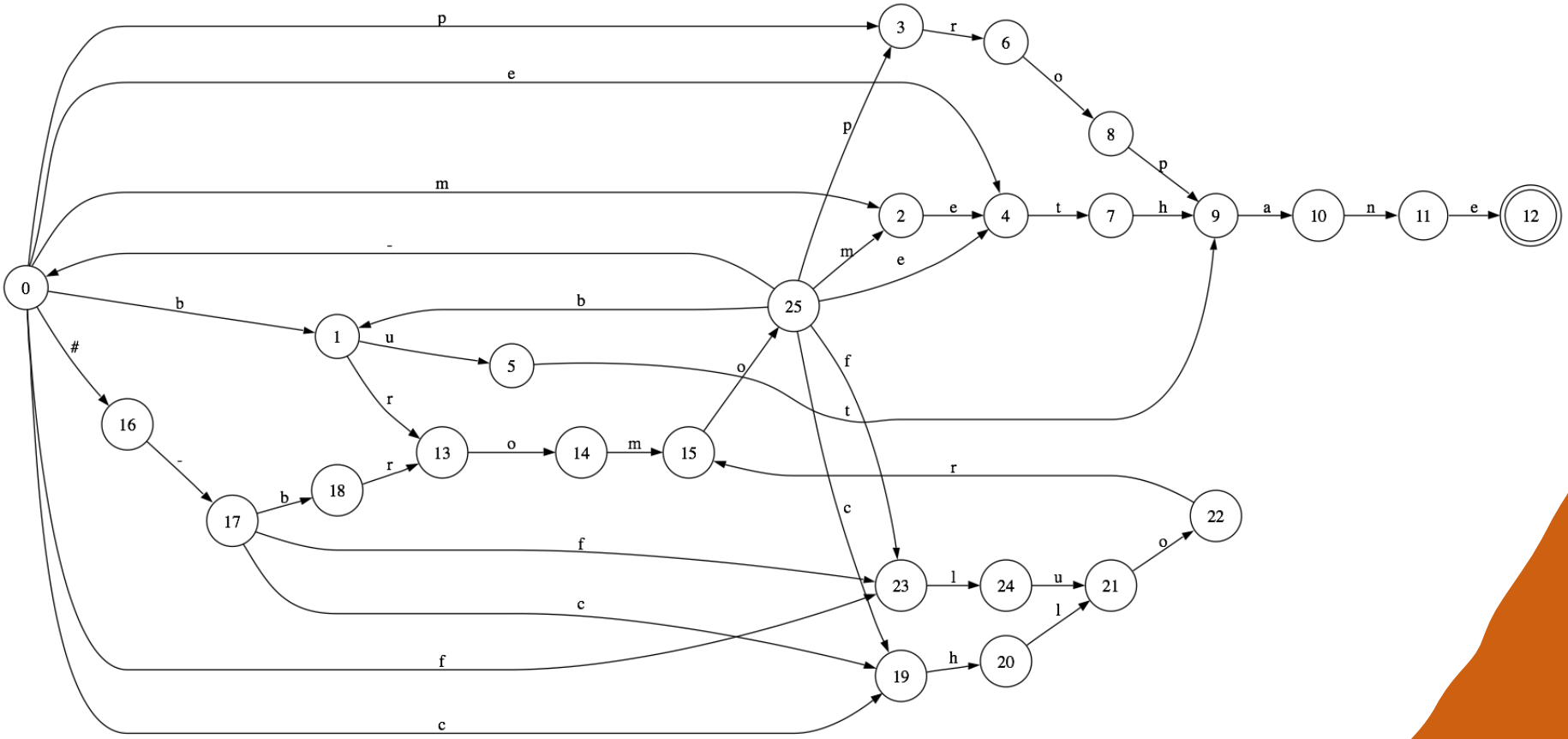
```
locant := “#”  /* any digit */

subst := “bromo” | “chloro” | “fluoro”
alk := “meth” | “eth” | “prop” | “but”
parent := alk “ane”

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

- 9-bromomethane
- 1-chloro-1-chloro-1-chloro-methane

REPRESENTING GRAMMARS AS DFAs



ENCODING TRIES, DAGS AND FSMS

- As described in the preceding slides, the nodes of the tries, DAGs and FSMs are of arbitrary degree (have variable fan-out).
- To reduce the complexity of this data structure one appraoch is to always allocate a size-of-alphabet array of out-going edges.
- A more efficient approach it to lower the data structure to binary nodes, containing “down” and “across” out-going edges.

CAFFEINEFIX DATA STRUCTURE

```
typedef struct {
    unsigned char ch;
    unsigned char state;
    unsigned int down;
    unsigned int across;
  } FSMType;
```

ALKANE EXAMPLE DECLARATION

```
static const FSMType fsm[15] = {
  /*  0 */ { 'b', 0,  3,  7 },
  /*  1 */ { 'r', 0,  2,  0 },
  /*  2 */ { 'o', 0,  9,  0 },
  /*  3 */ { 'u', 0, 11,  0 },
  /*  4 */ { 'h', 0,  5,  0 },
  /*  5 */ { 'a', 0,  6,  0 },
  /*  6 */ { 'n', 0, 13,  0 },
  /*  7 */ { 'e', 0, 14,  8 },
  /*  8 */ { 'm', 0, 10, 12 },
  /*  9 */ { 'p', 0,  5,  0 },
  /* 10 */ { 'e', 0, 14,  0 },
  /* 11 */ { 't', 0,  5,  0 },
  /* 12 */ { 'p', 0,  1,  0 },
  /* 13 */ { 'e', 1,  0,  0 },
  /* 14 */ { 't', 0,  4,  0 },
    };
```

CAFFEINEFIX PERFORMANCE

- As of March 2011, the actual CaffeineFix implementation contains 455,901 edges (trie states).
- The program to generate the table is about 8.1K lines of C, and takes 11 hours to run.
- Currently covers 231730/234142 (98.97%) of the names in NCI00;
  - c.f. Lexichem v2.0 231243/234142 (98.76%).
- 65109/71367 (91.23%) on Maybridge03
  - 68496 (95.98%) for Lexichem v2.0.

IMPROVED ENTITY SEGMENTATION

- In addition to full-name lookup, a convenient functionality is the ability to identify a string as a valid prefix.
- This is particularly useful in chemical NER, where valid IUPAC names can contain whitespace, parenthesis, brackets, hyphens, commas, periods, apostrophes and semicolons.
- Enables entity extension of “acetic”.

IS_PREFIX IMPLEMENTATION

```
bool is_prefix(const unsigned char *ptr) {
  unsigned int state = 0;

  for(;;) {
    if (fsm[state].ch == *ptr) {
      ptr++;
      if (*ptr == '\0')
        return true;  /* fsm[state].state */
      state = fsm[state].down;
    } else state = fsm[state].across;
    if (state == 0) return false;
  }
}
```

PUSH-DOWN AUTOMATA

- Unfortunately, DFAs are not powerful enough to capture the context-sensitive grammars needed for IUPAC-like names.
- The problem is nesting of parenthesis.
- Push-down automata are variants of DFAs that maintain an additional stack.
- This allows checking that parenthesis, brackets and braces are balanced and that open and close characters are matched.

SPELLING CORRECTION

- A relatively simple extension of the above exact match algorithm allows CaffeineFix's data structure to be used for automatic error correction.
- Backtracking allows consideration of substitution, insertion and deletion whilst traversing the finite state machine (FSM).
- Allows enumeration of all valid names within a specified edit-distance of a string.

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

AMBIGUITY IN NAME CORRECTION

- An interesting example of a potentially ambiguous name is “chlormethane”, which could be either “chloromethane” or “chlormethine” (ClCCN(C)CCCl).
- Another example is “sivastatin” that could be either “simvastatin” or “rivastatin”.
- Yet another is “carban” for which “carbon” or “carbane” could have been intended.

THE NEED FOR A WHITE WORD LIST

- “herein”, did you mean “heroin”.
- “aspiring”, did you mean “aspirin”.
- “cranium”, did you mean “uranium”.

- The “white word” dictionary currently contains 450 words and requires 1020 nodes, derived by spell correcting an English dictionary (/usr/dict/words).

OCR SPELLING CORRECTION

- In addition to Levenstein “string edit” distance, advantage can also be taken of the types of errors found in OCR text.
- Homoglyph substitutions such as between “1”, “l” and “I”, or “0” and “O”, or insertions of whitespace, hyphenation and line breaks (“
”) occur frequently.
- These can be penalized less than other edits/indels to improve spelling correction.

INFLUENCE OF GRAMMAR QUALITY

- The higher fidelity of the grammar, i.e. the higher the precision, the better the suggestions for spelling correction.

- “didec-2-ene” → “dodec-2-ene”
- “12-dichlorobenzene” → “1,2-dichlorobenzene”
- “spiro[2.2]hexane” → “spiro[2.3]hexane”
- “pyridine-8-carboxylic acid”

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

EXTRACT DICTIONARIES/FSMS

	Category	Words	Nodes	Fix
M	Molecule	Infinite	266,764	Y
D	Dictionary	11,196	24,985	Y
R	Registry #	Finite	211	N
C	CAS Number	Finite	4,779	N
E	Element	181	510	N
P	Fragment	Infinite	21,517	Y
A	Atom Fragment	11	46	N
Y	Polymer	25	201	N
G	Generic	115	385	N
N	Noise	15	61	N

HATTORI 2008 BENCHMARK SET

Set of 30 World and European Patents for U.S. Top selling drugs in 2005 described in

Kazunari Hattori, Hiroaki Wakabayashi and Kenta Tamaki, "**Predicting Key Example Compounds in Competitors' Patent Applications using Structural Information Alone**", *Journal of Chemical Information and Modeling*, Vol. 48, No. 1,pp. 135-142, 2008.

USPTO-50 DATA SET

US Patent #	Drug Name	M	D	Fix
US4231938	Mevacor	70	23	12	5
US4254129	Allegra	239	121	14	38
US4255431	Losec	198	66	23	11
US4282233	Claritin	100	25	12	7
US4335121	Flovent	231	84	17	20
US4382938	Ambien	78	23	3	8
US4503067	Coreg	255	132	10	9
US4572909	Norvasc	212	93	6	12
US4636505	Casodex	257	67	14	10
US4650884	Celexa	59	30	0	5
US4659516	Faslodex	478	208	18	49
US4659716	Clarinex	206	97	26	19
US4681893	Lipitor	174	64	14	17
US4689338	Aldara	409	293	7	54
US4695590	Zofran	88	21	32	7
US4721723	Paxil	89	33	11	5
US4738974	Nexium	83	29	20	2
US4746680	Meridia	342	182	15	20
US4755534	Lamisil	119	44	3	13
US4816470	Imitrex	253	77	24	9
US4847265	Plavix	68	25	7	4
US4874794	Abreva	47	31	7	14
US4895841	Aricept	327	157	18	37
US4935437	Arimidex	602	179	11	40
US4978672	Femara	393	146	28	34

US Patent #	Drug Name	M	D	Fix
US4990517	Vigamox	672	479	28	104
US5002953	Avandia	276	162	10	54
US5006528	Abilify	160	66	26	11
US5006530	Baycol	535	306	17	69
US5023269	Cymbalta	302	176	20	91
US5045552	Aciphex	412	291	20	135
US5116863	Patanol	288	130	22	44
US5153197	Cozaar	1095	746	51	149
US5196444	Atacand	104	54	10	3
US5270317	Avapro	369	186	21	27
US5354772	Lescol	308	143	20	29
US5360800	Lotronex	374	180	30	42
US5382600	Detrol	345	253	13	115
US5399578	Diovan	809	478	31	145
US5466823	Celebrex	744	479	11	68
US5521184	Gleevec	438	161	17	31
US5565447	Avodart	77	30	8	6
US5616599	Benicar	1451	725	16	206
US5633272	Bextra	706	307	33	52
US5747498	Tarceva	631	390	20	132
US5770599	Iressa	384	172	34	33
US5849911	Reyataz	712	343	52	104
US5859006	Cialis	329	126	14	27
US6410550	Chantix	293	141	25	17
US6566360	Levitra	372	173	15	33

CHEMICAL SYNONYMS IN PATENTS

Tradename	Count	Scientific Name	Count	Ratio
Levitra	474	Vardenafil	2610	5.51
Celebrex	5952	Celecoxib	16990	2.85
Aricept	2466	Donepezil	7164	2.91
Aldara	550	Imiquimod	4889	8.89
Cozaar	882	Losartan	9444	10.71
Benicar	176	Olmesartan medoxomil	2132	12.11
Detrol	579	Tolterodine	2812	4.86
Lescol	2261	Fluvastatin	14067	6.22
Casodex	2549	Bicalutamide	10979	4.31
Tarceva	3480	Erlotinib	5441	1.56
Patanol	124	Olopatadine	1223	9.86
Cialis	3562	Tadalafil	2105	0.59
Diovan	871	Valsartan	7196	8.26
Avapro	662	Irbesartan	6633	10.02
Flovent	484	Fluticasone propionate	11468	23.69
Bextra	1591	Valdecoxib	7450	4.68
Aciphex	331	Rabeprazole	2522	7.62
Lamisil	463	Terbinafine	6031	13.03
Vigamox	70	Moxifloxacin	2508	35.83
Femara	2044	Letrozole	8228	4.03
Zomig	221	Zolmitriptan	2840	12.85
Zofran	1021	Odansetron	1270	1.24
Spiriva	404	Tiotropium	5018	12.42
Coreg	639	Carvedilol	6050	9.47
Atacand	614	Candesartan	6408	10.44
Paxil	2027	Paroxetine	10981	5.42
Nasonex	214	Mometasone furoate	8275	38.67
Sustiva	2912	Efavirenz	6746	2.32
Arimidex	2376	Anastrozole	9973	4.20
Reyataz	558	Atazanavir	2084	3.73
			Avg	9.28

NAME CLASS FREQUENCY (USPTO-50)

	Category	Entities	Corrected
M	Molecule	8,947 (50.9%)	2,057 (11.7%)
D	Dictionary	916 (5.2%)	83 (0.5%)
R	Registry #	13 (0.1%)	
C	CAS Number	0 (0.0%)	
E	Element	1,369 (7.8%)	
P	Fragment	4,251 (24.2%)	
A	Atom Fragment	334 (1.9%)	
Y	Polymer	65 (0.4%)	
G	Generic	1,350 (7.7%)	
N	Noise	318 (1.8%)	
	Total	17,563	2,140 (12.2%)

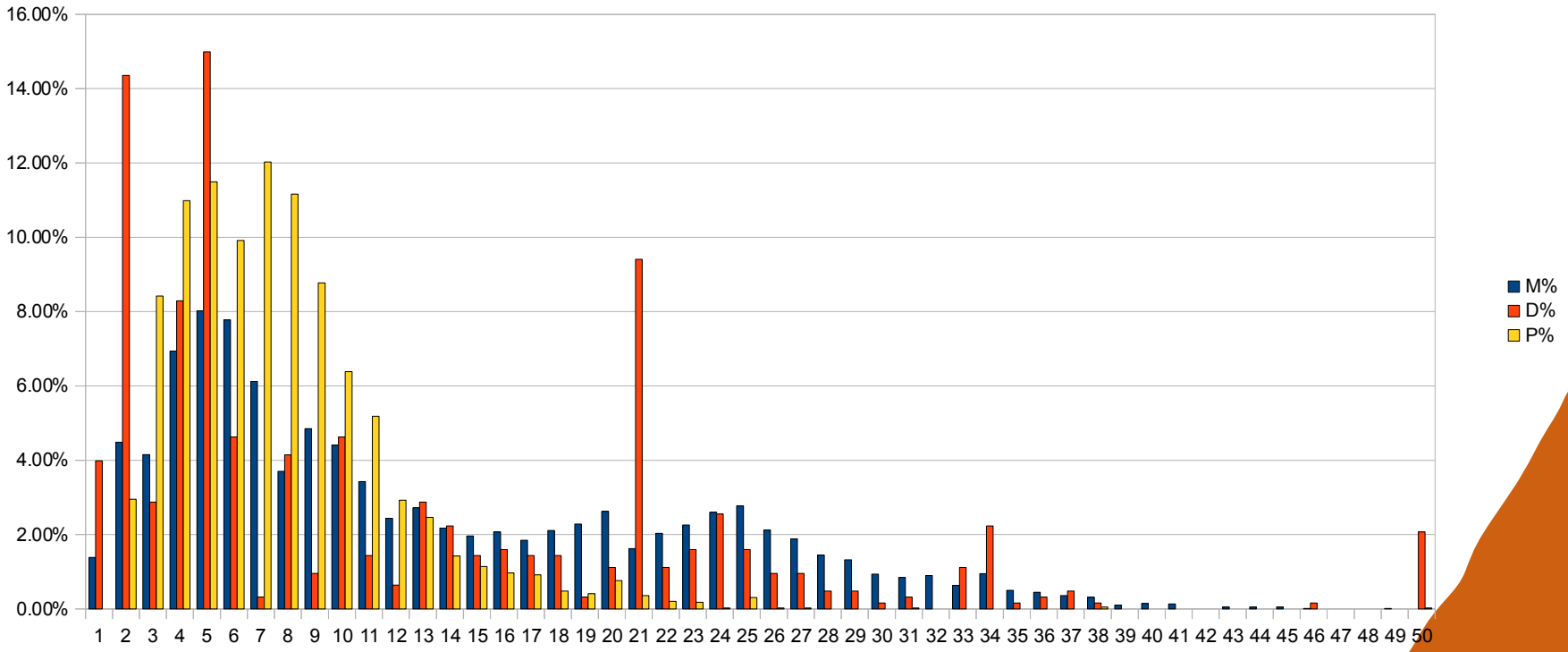
NAME-TO-STRUCTURE (USPTO-50)

	Category	Entities		SMILES	
M	Molecule	8,947	(50.9%)	7,583	(43.2%)
D	Dictionary	916	(5.2%)	627	(3.6%)
R	Registry #	13	(0.1%)	0	(0.0%)
C	CAS Number	0	(0.0%)	0	(0.0%)
E	Element	1,369	(7.8%)	1,305	(7.4%)
P	Fragment	4,251	(24.2%)	3,933	(22.4%)
A	Atom Fragment	334	(1.9%)	303	(1.7%)
Y	Polymer	65	(0.4%)	0	(0.0%)
G	Generic	1,350	(7.7%)	27	(0.2%)
N	Noise	318	(1.8%)	146	(0.8%)
	Total	17,563		13,924	(79.3%)

NAME CLASS FREQUENCY (IBM SIMPLE)

	Category	Entities	
M	Molecule	84,952,033	(33.8%)
D	Dictionary	27,352,706	(10.9%)
R	Registry #	313,196	(0.1%)
C	CAS Number	324,821	(0.2%)
E	Element	43,219,986	(18.6%)
P	Fragment	42,288,642	(18.0%)
A	Atom Fragment	4,363,901	(1.8%)
Y	Polymer	4,297,079	(0.9%)
G	Generic	26,905,972	(10.7%)
N	Noise	17,293,541	(6.9%)
	Total	251,311,877	

HEAVY-ATOM COUNT HISTOGRAMS



PATENT PART ANALYSIS (USPTO-50)

Part	Entities	'M'	'D'
Title	59	20	4
Abstract	403	35	7
Claims	1,738	486	13
Descr	15,292	8,383	876
Other	71	23	16
Total	17,563	8,947	916

PATENT PART ANALYSIS (IBM SIMPLE)

Part	Entities	
Title	1,415,735	(0.6%)
Abstract	7,239,915	(2.9%)
Claims	31,434,821	(12.5%)
Description	211,221,406	(84.0%)
Total	251,311,877	

CLASS/PART ANALYSIS (IBM SIMPLE)

	Title	Abstr	Claim	Descr
M	282605	1093287	7750118	75826023
D	111192	347804	317103	24617131
R	136	959	16345	295756
C	1	239	7478	317103
E	341962	2101469	6729482	34047073
P	114529	554683	5693225	35926205
A	13484	183747	962190	3204480
Y	16460	83262	479828	3717529
G	319652	1696578	5297828	19591914
N	215714	1177887	2221748	13678192

EXECUTIVE SUMMARY

- 251,311,877 chemical entities
- 8,336,922 patents with entities
- 5,025,487 unique names
- 3,329,224 unique SMILES
- 802,875 unique names not converted

HATTORI08 BENCHMARK LANGUAGES

Drug Name	Company	Patent	Language
Aciphex	Eisai Co	EP268956(A2)	EN
Aldara	3M Pharmaceuticals	EP145340(A2)	EN
Aricept	Eisai Co	EP296560(A2)	EN
Arimidex	AstraZeneca	EP296749(A1)	EN
Atacand	AstraZeneca	EP459136(A1)	EN
Avapro	Bristol-Myers Squibb	WO1991014679(A1)	FR
Benicar	Sankyo Pharma	EP503785(A1)	EN
Bextra	Pfizer	WO1996025405(A1)	EN
Casodex	AstraZeneca	EP100172(A1)	EN
Celebrex	Pfizer	WO1995015316(A1)	EN
Cialis	Lilly ICOS	WO1995019978(A1)	EN
Coreg	GlaxoSmithKline	DE2815926(A1)	DE
Cozaar	Merck & Co.	EP253310(A2)	EN
Detrol	Pfizer	EP0325571(A1)	EN
Diovan	Novartis	EP443983(A1)	DE
Femara	Novartis	EP236940(A2)	EN
Flovent	GlaxoSmithKline	NL8100707(A)	NL
Lamisil	Novartis	EP24587(A1)	EN
Lescol	Novartis	WO1984002131(A1)	EN
Levitra	Bayer	WO1999024433(A1)	DE
Nasonex	Schering-Plough	EP57401(A1)	EN
Patanol	Nestle SA	EP235796(A2)	EN
Paxil	GlaxoSmithKline	EP266574(A2)	EN
Reyataz	Bristol-Myers Squibb	WO1997040029(A1)	EN
Spiriva	Boehringer Ingelheim	EP418716(A1)	DE
Sustiva	Bristol-Myers Squibb	EP582455(A1)	EN
Tarceva	Genentech	WO1996030347(A1)	EN
Vigamox	Alcon	EP550903(A1)	DE
Zofran	GlaxoSmithKline	DE3502508(A1)	DE
Zomig	Medpointe Pharm	WO1991018897(A1)	EN

IBM SIMPLE PATENT DATABASE ABSTRACT LANGUAGES

Code	Language	Count	Fraction
en	English	9584057	79.98%
fr	French	1753943	14.64%
de	German	610485	5.09%
ja	Japanese	15467	0.13%
es	Spanish	11157	0.09%
zh	Chinese	3770	0.03%
ko	Korean	3464	0.03%
ru	Russian	392	0.00%
fi	Finnish	206	0.00%
pt	Portuguese	146	0.00%
nl	Dutch	136	0.00%
sv	Swedish	122	0.00%
no	Norwegian	59	0.00%
da	Danish	10	0.00%
it	Italian	10	0.00%
tr	Turkish	6	0.00%
ar	Arabic	1	0.00%
el	Greek	1	0.00%
sl	Slovene	1	0.00%
		11983434	100.00%

LANGUAGE TRANSLATION REFERENCE

- Roger Sayle, "**Foreign Language Translation of Chemical Nomenclature by Computer**", *Journal of Chemical Information and Modeling*, Vol. 49, No. 3, pp. 519-530, 2009.

JAPANESE, CHINESE AND RUSSIAN

- 119/9266 (1.28%) AZ-interest patents are Japanese, 0.74% German, 0.49% French.
- Example JP2007176809A has 52 names.
- NextMove's “extract” program makes use of OpenEye's Lexichem for chemical nomenclature translation.
- This is performed as a pre-processing step to generate partially translated XML that is then text mined for English entities.

EXAMPLE OF NAME TRANSLATION

- Chinese (from CN101622315A)

2,4-双(2-羟基-4-丙氧基-苯基)-6-(2,4-二甲基苯基)-1,3,5-三嗪

- English translation

2,4-bis(2-hydroxy-4-propoxy-phenyl)-6-(2,4-dimethylphenyl)-1,3,5-triazine

- Extracted SMILES

CCCOc1ccc(c(c1)O)c2nc(nc(n2)c3ccc(cc3O)OCCC)c4ccc(cc4C)C

AND BECAUSE WE'RE IN GOSLAR...

From the Atacand Patent, EP45136(B)

- Patentanspruche

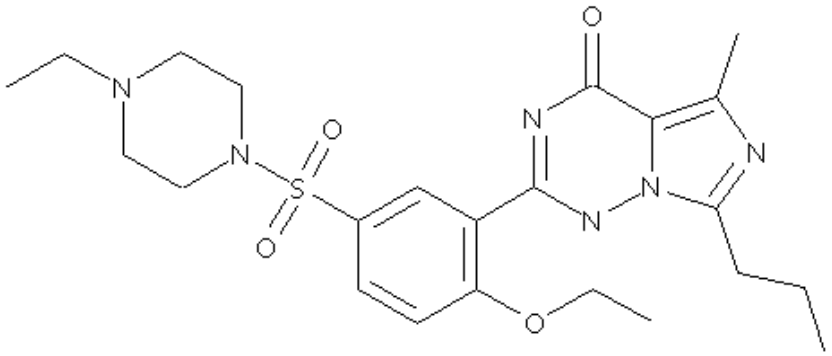
  1. 2-Ethoxy-1-[[2'-(1 H-tetrazol-5-yl)biphenyl-4-yl]methyl]benzimidazol-7-carbonsaure oder ein pharmazeutisch annehmbares Salz davon.

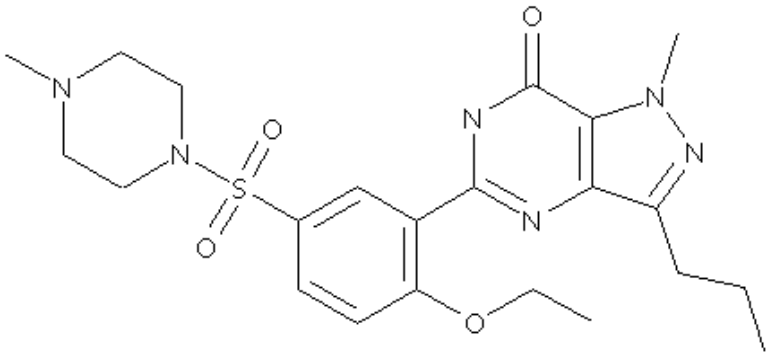
- 2-ethoxy-1-[[2'-(1H-tetrazol-5-yl)biphenyl-4-yl]methyl]benzimidazole-7-carboxylic acid
- CCOc1nc2cccc(c2n1Cc3ccc(cc3)c4ccccc4c5[nH]nnn5)C(=O)O

SIMILARITY RESULTS (MACCS KEYS)

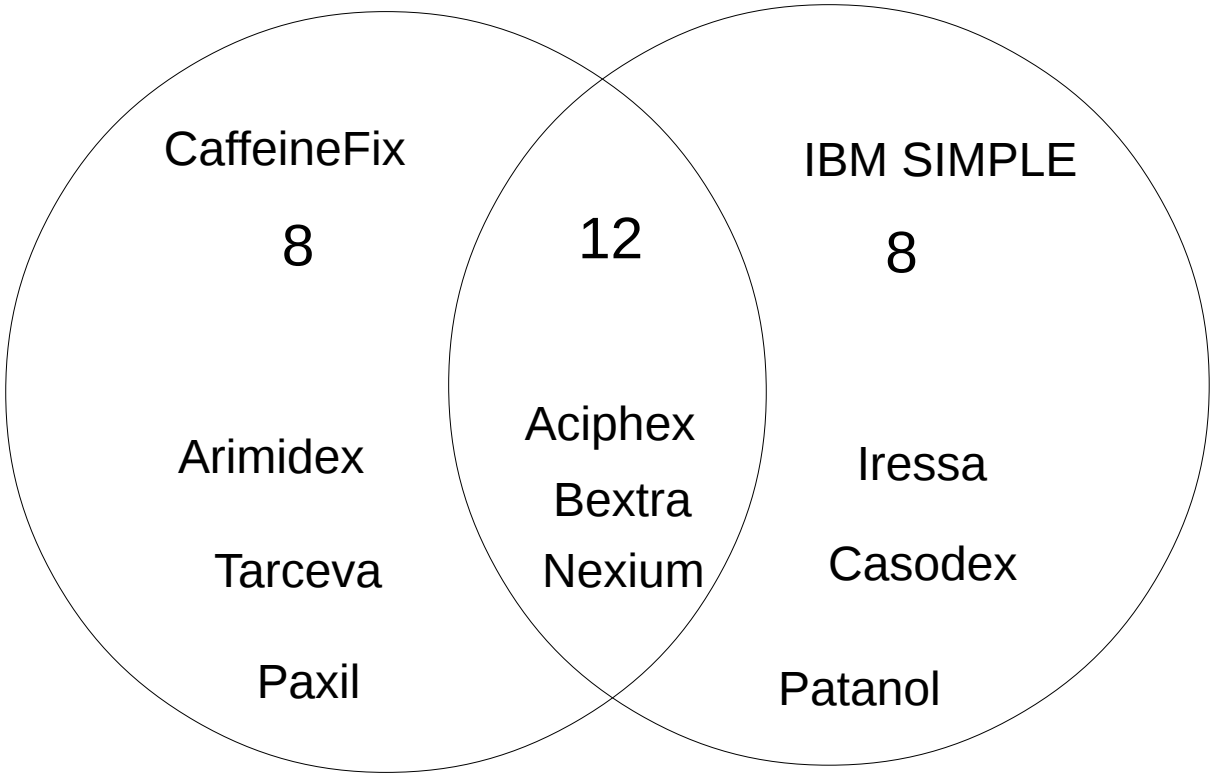
1.00	Abilify,Abreva, Aciphex, Aldara, Arimidex, Atacand, Bextra, Casodex, Celebrex, Cozaar, Cymbalta, Detrol, Imitrex, Losec, Meridia, Mevacor, Nexium, Norvasc, Paxil, Tarceva (20)
0.95-1.00	Diovan, Levitra
0.90-0.95	Avandia, Baycol, Lipitor, Lotronex
0.85-0.90	Allegra, Benicar, Celexa, Femara
0.80-0.85	Aricept, Avapro, Iressa, Lamisil, Vigamox
0.75-0.80	Ambien, Zofran
0.70-0.75	Cialis, Claritin, Coreg, Flovent, Gleevec, Lescol, Reyataz
0.65-0.70	Chantix, Clarinex, Plavix
0.60-0.65	Avodart, Faslodex, Patanol

EXAMPLE NEAR MISS (LEVITRA)





VENN DIAGRAM OF SYNERGIES



22

Ambien

Cialis

Claritin

Flovent

Zofran

PERFORMANCE

	MaxErr	CSV	SMI	USMI	HITS	AVGT	CPU
CaffeineFix	None	15,687	12,672	4534	14	0.82	3.7s
CaffeineFix	0	16,850	13,534	4903	18	0.87	5.5s
CaffeineFix	1	19,307	15,014	6066	20	0.88	130s
CaffeineFix	2	21,250	16,172	6416	21	0.89	163m36
SIMPLE	None			6399	20	0.90	
Both	None			9039	23	0.91	
Both	0			9258	27	0.93	
Both	1			10,104	28	0.93	

IMPROVEMENT EXAMPLES

- No Spelling Correction
  - Abilify, Abreva, Aldara, Bextra, Celebrex, Cymbalta, Detrol, Imitrex, Losec, Meridia, Mevacor, Nexium, Paxil
- MaxDist = 0
  - Arimidex, Atacand, Cozaar, Norvasc
- MaxDist = 1
  - Aciphex*, Tarceva

INFORMATIVE FAILURES

- Aricept
  - ...-4-((5,6-dimethoxy-1-indanon)-2-yl)propylpiperidine
- Avandia
  - ...-2,4-thiazolidinedione
- Celexa
  - (4'-fluorophenyl)-1,3-dihydroisobenzofuran-5-carbonitrile
- Iressa
  - 4-(3'-chloro-4'-fluoroanilino)-...

INTERESTING FACTS

- The name “cyclosporin A” appears 15,639 times in a single patent, US20070015693.

- The SMILES NCCO is found as 93 different synonyms in patents
  - 2-amino-ethanol, 2-hydroxyethylamine, 1-amino-2-ethanol, 2-amino-1-ethanol, 2-hydroxyethylamine, 1-amino-ethan-2-ol, 2-amino ethy alcohol, (2-hydroxyethyl)amine,...

CONCLUSIONS

- A chemical text mining method incorporating grammars (infinite dictionaries), language translation and automatic spelling correction has been presented.
- The use of IUPAC and IUPAC-like grammars is shown to have significant benefits over simple dictionary-based methods for patent analysis.
- Both language translation and automatic spelling correction are also shown to be advantageous.

FUTURE WORK

- Next generation spelling correction
  - To handle high string-edit-distance issues.
- Frequency-dependent disambiguation
  - More common tokens are better suggestions.
- Context-dependent disambiguation
  - Take advantage of surrounding names.

ACKNOWLEDGEMENTS

- AstraZeneca R&D, Sweden.
- Steve Boyer, IBM/Collabra, USA.
- Pat Walters, Vertex, USA.
- Daniel Lowe, University of Cambridge, UK.
- Nicko Goncheroff, SureChem, UK.
- Alexander “Sandy” Lawson, Elsevier, DE.