Translating IUPAC-like Chemical Nomenclature to and from Simplified Chinese

Roger Sayle and Daniel Lowe

NextMove Software

Cambridge, UK

ACS National Meeting, Philadelphia, USA 22nd August 2012

MOTIVATION

- Whilst most scientific articles are written in English, a significant fraction of chemistry research remains only available in other languages.
- The increasing use of overseas contract research organizations is increasing the prevalence of foreign languages in Electronic Lab Notebooks and internal technical reports.
- Existing language tools, such as Google translate, often perform poorly on technical (domain specific) terminology, such as IUPAC-like nomenclature.

GOOGLE SEARCH STATISTICS

Lang	Query	Search Hits
en	benzoic acid	6,670,000
zh	苯甲酸	4,090,000
ja	安息香酸	2,000,000
es	ácido benzoico	439,000
ru	Бензойная кислота	310,000
de	benzoesäure	267,000
fr	acide benzoïque	179,000
nl	benzoëzuur	139,000

IBM SIMPLE PATENT DATABASE C07D ABSTRACT LANGUAGES

Code	Language	Count	Fraction
en	English	316338	70.96%
fr	French	104924	23.54%
de	German	22694	5.09%
ja	Japanese	1178	0.26%
zh	Chinese	310	0.07%
es	Spanish	230	0.05%
ko	Korean	120	0.03%
ru	Russian	16	0.00%
pt	Portuguese	7	0.00%
fi	Finnish	4	0.00%
		445821	100.00%

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

ACS National Meeting, Philadelphia, USA 22nd August 2012

PREVIOUS WORK

- Roger Sayle, "**Foreign Language Translation of Chemical Nomenclature by Computer**", *Journal of Chemical Information and Modeling*, Vol. 49, No. 3, pp. 519-530, 2009.
- Roger Sayle, "**Preserving Nuance in Chemical Nomenclature Translation**", *The ATA Chronicle* (The Journal of the American Translators Association), Vol. 38, October 2009.

INTERNATIONAL NOMENCLATURE

- Morphology/semantics preserved across languages:
  - English: 2-(4-chlorophenoxy)acetic acid
  - Chinese: 2-(4-氯苯氧基)醋酸
  - German: 2-(4-chlorphenoxy)essigsäure
  - French: acide 2-(4-chlorophénoxy)acétique
  - Spanish: ácido 2-(4-clorofenoxi)acético
  - Swedish: 2-(4-klorofenoxi)ättiksyra
  - Italian: acido 2-(4-clorofenossi)acetico
  - Polish: kwas 2-(4-chlorofenoksy)octowy
  - Japanese: 2-(4-クロロフェノキシ)酢酸

CHINESE ELEMENTS

- hydrogen 氢 [H]
- oxygen 氧 [O]
- nitrogen 氮 [N]
- gold 金 [Au]
- silver 银 [Ag]
- iron 铁 [Fe]
- lead 铅 [Pb]

CHINESE ALKANES

• methane	甲烷	C
• ethane	乙烷	CC
• propane	丙烷	CCC
• butane	丁烷	CCCC
• pentane	戊烷	CCCCC
• hexane	己烷	CCCCCC
• heptane	庚烷	CCCCCCC

CHINESE ALKYL GROUPS

• methyl	甲基	*C
• ethyl	乙基	*CC
• propyl	丙基	*CCC
• butyl	丁基	*CCCC
• pentyl	戊基	*CCCCC
• hexyl	己基	*CCCCCC
• heptyl	庚基	*CCCCCCC

CHINESE FUNCTIONAL GROUPS

• hydroxy-	羟基	*O
• amino-	胺基	*N
• fluoro-	氟	*F
• chloro-	氯	*Cl
• bromo-	溴	*Br
• iodo-	碘	*I
• nitro-	硝基	*N(=O)=O

CHINESE CARBOCYCLES

• benzene	苯	c1ccccc1
• naphthalene	萘	c1ccc2ccccc2c1
• azulene	薁	
• anthracene	蒽	
• phenanthrene	菲	
• fluorene	芴	
• pyrene	芘	

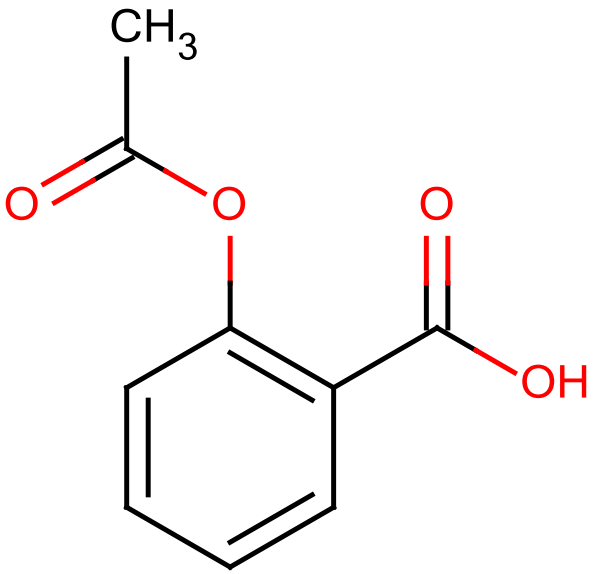
CHINESE HETEROCYCLES

• furan	呋喃	o1cccc1
• pyrrole	吡咯	[nH]1cccc1
• thiophene	噻吩	s1cccc1
• pyridine	吡啶	n1ccccc1
• pyridazine	哒嗪	n1nccccc1
• pyrimidine	嘧啶	n1cnccc1
• pyrazine	吡嗪	n1ccncc1

CHINESE CARBOXYLIC ACIDS

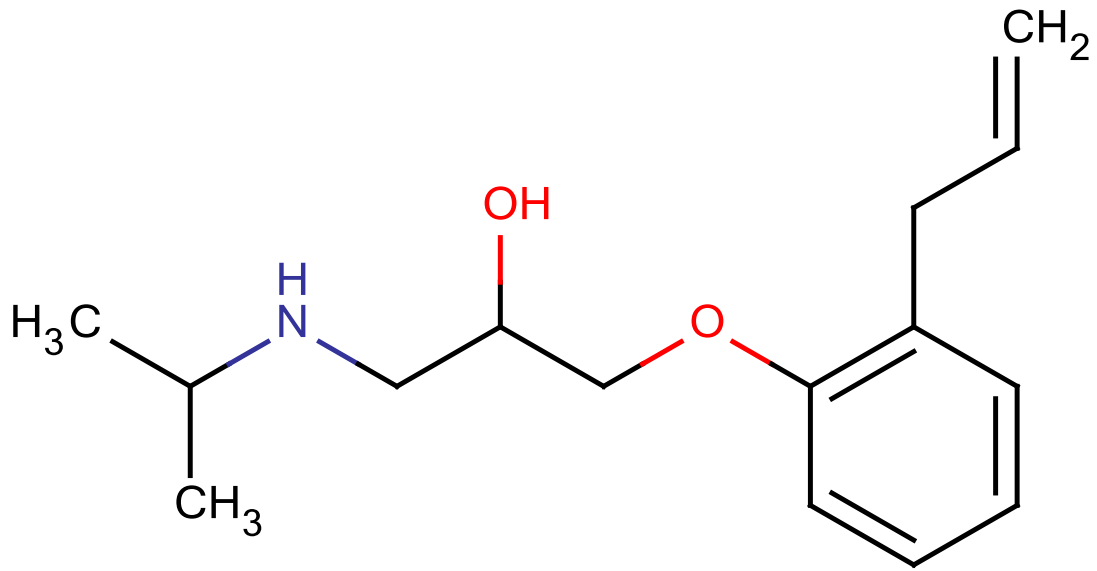
- formic acid 蚁酸 C(=O)O
- acetic acid 醋酸 CC(=O)O
- propionic acid 丙酸 CCC(=O)O
- butyric acid 酪酸 CCCC(=O)O
- valeric acid 戊酸 CCCCC(=O)O
- hexanoic acid 己酸 CCCCCC(=O)O
- enanthic acid 庚酸 CCCCCCC(=O)O

EXAMPLE #1: ASPIRIN



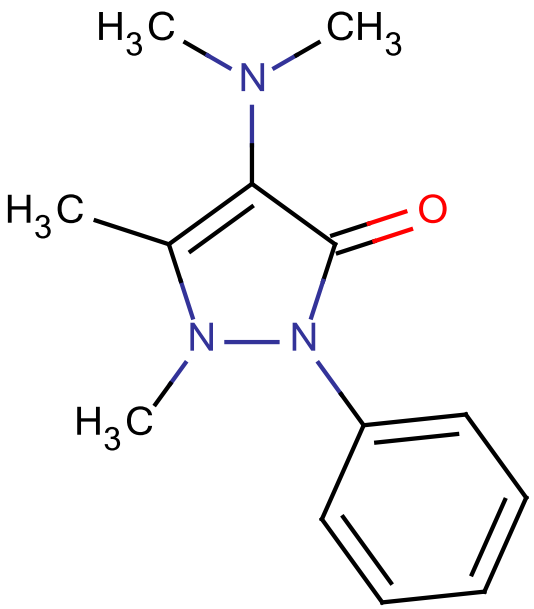
- 2-acetoxybenzoic acid
- 2-乙酰氧基苯甲酸

EXAMPLE #2: ALPRENOLOL



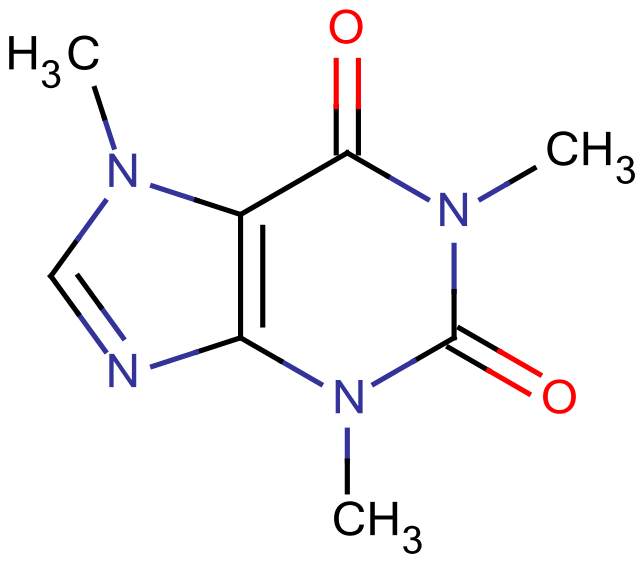
- 1-(2-allylphenoxy)-3-(isopropylamino)propan-2-ol
- 1-(2-烯丙基苯氧基)-3-(异丙基氨基)丙-2-醇

EXAMPLE #3: AMINOPYRINE



- 4-dimethylamino-1,5-dimethyl-2-phenyl-pyrazol-3-one
- 4-二甲氨基-1,5-二甲基-2-苯基–吡唑-3-酮

EXAMPLE #4: CAFFEINE



- 1,3,7-trimethylpurine-2,6-dione
- 1,3,7-三甲基嘌呤-2,6-二酮

SIMPLIFIED CHINESE TO ENGLISH FOR CHEMICAL TEXT MINING OF PATENTS

- Early efforts in (Chinese) chemical nomenclature translation focused on translating and round tripping delimited and often machine-generated IUPAC names.
- More recently focus has shifted to the application of this technology to performing chemical named entity recognition (CNER) in pharmaceutical patents.
- The rest of this presentation describes the current state-of-the-art on this challenge, and the influence of Chinese OCR on compound retrieval rates.

LEADMINE INTRODUCTION

- LeadMine is NextMove Software's application for chemical text mining, built upon CaffeineFix for automatic chemical spelling correction.
- R. Sayle, P.H. Xie and S. Muresan, "Improved Chemical Text Mining of Patents with Infinite Dictionaries and Automatic Spelling Correction", J. Chem. Inf. Model. 52(1), pp. 51-62, January 2012.
- Recent improvements to LeadMine and CaffeineFix were described earlier in the week in the CINF session on "Chemical Information in Patents".

LEADMINE V2.0 CHINESE SUPPORT

- Chinese is supported by preprocessing the input text, substituting/transliterating the hanzi characters found in IUPAC-like names into English.
- The current translation dictionary (June 2012) includes 1046 substitution rules.
- Traditional Chinese is supported by internally automatically mapping 166 traditional characters found in IUPAC-like names to their simplified forms.
- Efficient matching is made possible by “compiling” these rules into a 7000 line Java source file.

EXAMPLE NOISY INPUT

5.如权利要求1所述的化合物，所述化合物是：2_(4-氯-苯氧基)-2-甲基-1^-噻唑-2-基-丙酰胺；2—甲基-2-(4-甲硫基-苯氧基)-N-噻唑-2-基-丙酰胺;2-(6-氯-吡啶-2-基氧基)-2-甲基-N-噻唑-2-基-丙酰胺； 2_甲基-2-(萘-1 -基氧基)-N-噻唑-2-基-丙酰胺； 2_甲基-2-(萘-2-基氧基)->^噻唑-2-基-丙酰胺； 2-(2，4-二氟-苯氧基)-2-甲基-N-噻唑-2-基- 丙酰胺； 2-(4-氟-苯硫基)-2-甲基-N-痤唑-2-基-丙酰胺； 2-甲基-2-(4-苯氧基-苯氧基)-1^-噻唑-2-基-丙酰胺； 2_甲基-N-噻唑-2-基-2-(4'-三氟甲氧基-联苯-4-基氧基)-丙酰胺； 2-(苯并[1，3]二氧杂环戊烯-5-基氧基)-^(5-氯-噻唑-2-基)-2-甲基-丙酰胺；N-(5-氯-噻唑-2-基)-2-(2，4-二氟-苯氧基)-2-甲基-丙酰胺； 2-(5-氯-吡啶-3-基氧基)-N-(5-氯-痤唑-2-基)-2-甲基-丙酰胺；...

EXAMPLE ANNOTATED OUTPUT

<CLM>5.如权利要求1所述的化合物，所述化合物是：
2_(4-chloro-phenoxy)-2-methyl-1^-thiazol-2-yl-propanamide;
2—methyl-2-(4-methylthio-phenoxy)-N-thiazol-2-yl-propanamide;
2-(6-chloro-pyridin-2-yloxy)-2-methyl-N-thiazol-2-yl-propanamide;
2_methyl-2-(naphthalene-1 -yloxy)-N-thiazol-2-yl-propanamide;
2_methyl-2-(naphthalen-2-yloxy)->^thiazol-2-yl-propanamide;
2-(2，4-difluoro-phenoxy)-2-methyl-N-thiazol-2-yl- propanamide;
2-(4-fluoro-benzenethio)-2-methyl-N-瘗azol-2-yl-propanamide;
2-methyl-2-(4-phenoxy-phenoxy)-1^-thiazol-2-yl-propanamide;
2_methyl-N-thiazol-2-yl-2-(4'-trifluoromethoxy-biphenyl-4-yloxy)-propanamide;
2-(benzo[1，3]dioxacyclopenten-5-yloxy)-^(5-chloro-thiazol-2-yl)-2-methyl-propanamide;
N-(5-chloro-thiazol-2-yl)-2-(2，4-difluoro-phenoxy)-2-methyl-propanamide;
2-(5-chloro-pyridin-3-yloxy)-N-(5-chloro-瘗azol-2-yl)-2-methyl-propanamide;
N-(5-chloro-thiazol-2-yl)-2-methyl-2-(3-nitro-phenoxy)-propanamide;

BENCHMARK #1: CN101622231A

- The UTF-16 encoded XML for this patent was kindly provided/suggested by Steve Boyer (IBM/Collabra).
- This document corresponds to US patent application 2010/0144772 A1, the non-OCR full text is available from Google patents.
- Claim #5 lists 168 fully exemplified structures explicitly claimed by this patent.

BENCHMARK #1: CN101622231A

- 144/168 (85.7%) can be recognized by LeadMine v2 from the English text (US 2010144772).
- 0/168 (0%) can be recognized by LeadMine from the text translated using Google translate.
- 5/168 (3%) could be recognized by LeadMine v1, which uses OpenEye Scientific Software's Lexichem for Chinese translation.
- 71/168 (42%) can be recognized by LeadMine v2, which uses NextMove Software's Chinese translation.
- Only 9/168 (5.4%) when OCR correction is disabled.

BENCHMARK #1: CN101622231A

- Three compounds (39, 79 and 161) can be recovered from the Chinese text but could not be found in the English patent! These appear to be typos corrected during translation.
- A mistake by the translators has also resulted in an accidental duplication of names 33, 34 and 35, such that there are 171 compounds in the CN patent.
- A significant cause of failures is the incorrect OCR of the “甲” as found in methyl, which is frequently incorrectly perceived as “甲”.

BENCHMARK #2: CN1964970

- The non-searchable PDF of CN1964970 was provided by the NIBR-IT text mining group at Novartis.
- These images can be freely downloaded via EPO.
- This document corresponds to US patent application US20050234042.
- The challenge is to extract the 57 exemplified structures from claim #24 (on pages 19-23).
- This benchmark tested the quality/impact of Chinese Optical Character Recognition (OCR) software.

BENCHMARK #2: CN1964970

- Two OCR approaches were evaluated; the open source tesseract from HP/Google and the commercial product Abbyy FineReader version 11.
- For tesseract, conversion to 300x300 dpi image files was done using GPL Ghostscript v9.05.
- Of the two lossless image file formats supported by tesseract's leptonica library (v1.68), TIFF g3 format produced smaller files than PNG.
- Abbyy FineReader processes pages significantly faster than the 1 min/page of Google's tesseract.

BENCHMARK #2: CN1964970

- 50/57 (88%) of compounds can be found by LeadMine from the English patent, US20050234042.
- 0/57 compounds can be found using Abbyy FineReader's pure "Simplified Chinese" setting.
- 4/57 (7%) can be found using FineReader's hybrid "Simplified Chinese and English" language setting.
- 2/57 (3.5%) can be found using HP/Google's open source tesseract package.
- Similar results were also seen with CN1684951.

CONCLUSIONS

- In some respects, the ability to mine any chemical structures from Chinese patents is encouraging.
- On high quality OCR input, about 50% of the names found in English CNER can be retrieved from Chinese.
- Results are highly dependent upon the quality of OCR software used (IBM looks to have/use the best).
- Abbyy FineReader's "Simplified Chinese and English" performed best of the tested OCR software.
- Interestingly, some things are only found in Chinese.

POTENTIAL FUTURE WORK

- Evaluate additional Chinese OCR packages, include Adobe Acrobat, COCR2, ReadIris 14 and Dan Ching.
- Teach LeadMine's Chinese translation about more problematic homoglyph errors from Chinese OCR.
- Make use of (tesseract's) OCR per character confidence during CaffeineFix spelling correction.
- Perhaps fund a (Chinese speaking) summer student to develop improved "training data" files for Google's tesseract, to better support the fonts used by the Chinese patent office.

ACKNOWLEDGEMENTS

- Sorel Muresan, Plamen Petrov and Paul Hongxing Xie, AstraZeneca R&D, Molndal, SE.
- Steve Boyer, IBM/Collabra, San Jose, USA.
- Fatma Oezdemir-Zaech, NBIR-IT, Novartis, Basel, CH.
- OpenEye Scientific Software, Santa Fe, USA.
- Wikipedia.

- Thank you for your time.