CHEMISTRY ENABLING CHINESE, JAPANESE AND KOREAN PATENTS

Daniel Lowe and Roger Sayle, NextMove Software Ltd, Cambridge, UK
daniel@nextmovesoftware.com

Introduction

Chinese, Japanese and Korean (CJK) patents account for over half of all national patent filings and hence are of increasing importance to patent informatics. In the domain of chemistry the specific chemicals mentioned in a patent are often crucial for finding relevant patents and prior art searching.

The most interesting chemicals in a patent are typically the novel ones, which are described using systematic chemical nomenclature. Fortunately while CJK chemical names appear quite different to European chemical names, the underlying nomenclature is essentially the same. Hence we have developed tools to translate CJK chemical names to English such that existing chemical entity recognition and conversion tools may be used.

Asian chemical nomenclature

6-aminopyrimidine-2,4,5-triol

Chinese (Hanzi used for each morpheme)

6-氨基嘧啶-2,4,5-三醇

ammonia radical | pyrimidine | three | alcohol

Japanese (Phonetic translation to Katakana)

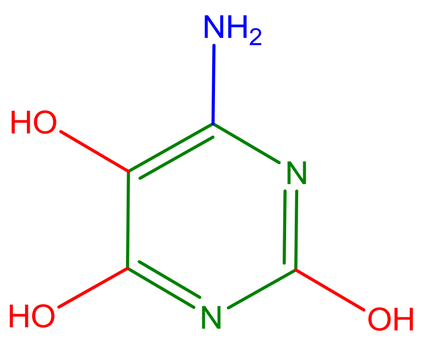
6-アミノピリミジン-2,4,5-トリオール

amino | pyrimidine | tri | ol

Korean (Phonetic translation to Hangul)

6-아미노피리미딘-2,4,5-트리올

amino | pyrimidine | tri | ol



Challenging cases

Word order (Chinese)

Chinese	Literal translation	English
间硝基氯化苄	meta-nitrochlorine of benzyl	meta-nitrobenzyl chloride

Context sensitivity (Chinese)

酮 ketone	环己**酮** cyclohexan**one**	2-**酮**丁酸 2-**keto**butanoic acid

Different nomenclature used in English (Chinese)

Chinese	Literal translation	English
丁酮二酸	butanonedioic acid	oxaloacetic acid

Implicit ester word (Japanese/Korean)

Japanese	Literal translation	English
酢酸エチル	acetic acid ethyl	acetic acid ethyl ester

Adding spaces in English translation (Chinese/Japanese)

Japanese	Literal translation	English
ピペリジン臭化水素酸塩	piperidinehydrobromide	piperidine hydrobromide

Acid character has two interpretations (Chinese/Japanese/Korean)

醋**酸** acet***ic acid***	醋**酸**钠 acet***ate*** sodium

OCR errors

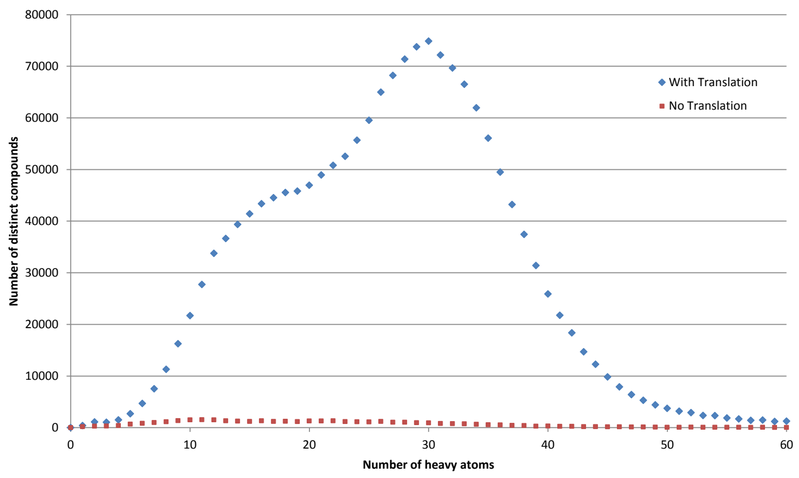
Patent text	OCR error	Translation after fix
己二酸二甲酯	己 should be 己	hexanedioic acid dimethyl ester

Features

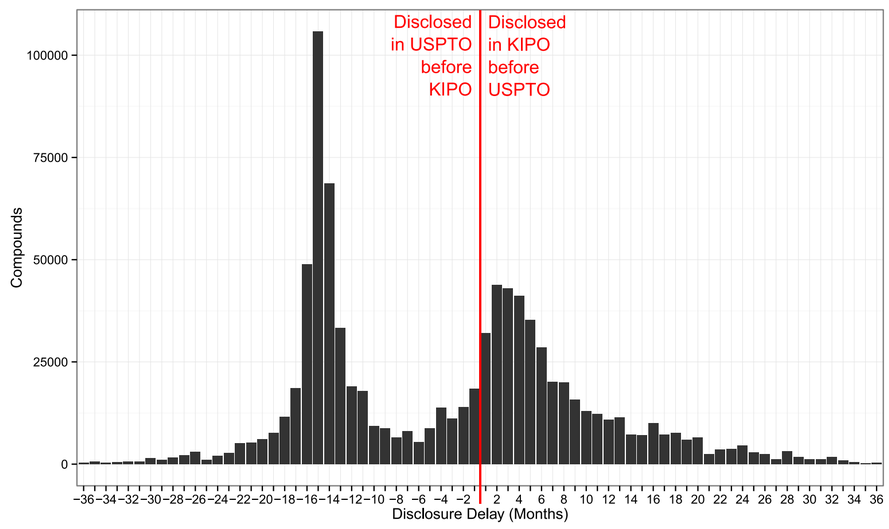
- Simplified and traditional Chinese characters are accepted interchangeably where their meaning is equivalent
- Erroneous spaces and common OCR errors handled in Chinese documents
- KIPO (Korean Intellectual Property Office) patents as PDFs can be converted to XML with patent sections annotated
- The translation tool is available as a Java library or command-line utility

Example application: Database of chemicals from Korean patents

63 thousand Korean patent applications (as PDF) were analysed (spanning 1990 to March 2015). After translation, LeadMine[†] was used to extract 1,740,040 distinct compounds with a mean heavy atom count of 27.1. By contrast without translation only 39,824 distinct compounds could be extracted!



Compared to an analysis of USPTO patents (1976-March 2015) 230,770 compounds were novel to the Korean patents. In the period 2006-2014, 46% of compounds appeared in a KIPO filing before a USPTO filing.



Performance on Japanese and Chinese

DataSet	Same structure obtained from English name and translated Asian name*
51,215 Chinese/English catalogue names (from chemBlink)	87.1%
260,965 Japanese/English catalogue names (from ChemicalBook)	88.3%

*As the Asian name may be semantically different to the English name (due to a mistake in the catalogue) this gives a lower bound on actual performance

Conclusions

We have developed state of the art chemical name translation for Chinese, Japanese and Korean chemical names. As an example, we have shown that over a million chemical compounds can be extracted from Korean patents and that many of these are novel and/or published earlier than their USPTO counterparts.

[†] Lowe, D. M.; Sayle, R. A. LeadMine: A Grammar and Dictionary Driven Approach to Entity Recognition. *J. Cheminformatics* **2015**, *7*, S5.