Legal Information Management, 12 (2012), pp. 173–184

doi:10.1017/S1472669612000436

Searching Legal Information in Multiple Asian Languages

Abstract: In this article Philip Chung, Andrew Mowbray, and Graham Greenleaf, the Co-Directors of the Australasian Legal Information Institute (AustLII), explain the need for an open source search engine which can search simultaneously over legal materials in European languages and also in Asian languages, particularly those that require a 'double byte' representation, and the difficulties this task presents. A solution is proposed; the 'u16a' modifications to AustLII's open source search engine (Sino) which is used by many legal information institutes. Two implementations of the Sino u16A approach, on the Hong Kong Legal Information Institute (HKLII), for English and Chinese, and on the Asian Legal Information Institute (AsianLII), for multiple Asian languages, are described. The implementations have been successful, though many challenges (discussed briefly) remain before this approach will provide a full multi-lingual search facility.
Keywords: free legal information; information retrieval; search engines; Asian languages

INTRODUCTION

The open source search engine, Sino[1], developed by the Australasian Legal Information Institute (AustLII), is used by a significant proportion of the current free access Legal Information Institutes (LIIs),[2] and by their shared portal, the World Legal Information Institute (WorldLII) (Greenleaf, 2012). AustLII is a free access LII which has operated since 1995, the second to be established, and now provides over 500 databases of Australasian legal materials (Greenleaf, Mowbray and Chung, 2011, 2010). The Sino search engine, and other software developed for large-scale development of legal information systems, such as hypertext mark-up software and the LawCite citator software, have been developed by AustLII's Directors and technical staff, in stages since the early 1990s. Over the last decade they have been provided by AustLII, with technical support, to other free access LIIs (Greenleaf, 2012; Greenleaf, Mowbray and Chung, 2011, 2010).

Sino is also used by the Asian Legal Information Institute (AsianLII – www.asianlii.org), a non-profit and free access website for legal information from all 28 countries and territories with separate legal jurisdictions in Asia.[3] Its coverage is from Japan in the east, to Pakistan in the west, and from Mongolia in the north, to Timor Leste in the south. AsianLII has been available for public access since December 2006. It provides for searching and browsing over 300 databases of legislation, case law, law reform reports, law journals and other legal information, where available, from each country in the region. All databases can be searched simultaneously, or searches can be limited to one country's databases or other combinations. Search results can be ordered by relevance, by date, or by database (Greenleaf, Chung and Mowbray, 2008).

There are many reasons why free access to law providers wish to use open source software, including 'avoidance of monopolistic control, reuse of information, standardization, tools sharing, avoidance of revenue models depending on selling information as a product' (Poulin, Mowbray and Lemyre, 2007). At least twenty open source search engines are available from which LIIs can choose, if they do not wish to purchase a proprietary search engine.[4] However, none have been adopted by free access law providers to the extent that Sino has. Sino is available to be used by LIIs in any part of the world, and this may involve a need for the searching of legal information in any of the world's languages. Its existing user base, the advantages of its global usability by any free access to law provider, and its use on a global legal portal, all make it desirable that Sino should have a very broad multi-lingual search capacity. However, the original version of Sino was effective in searching texts only in European languages and other languages (such as Bahasa Indonesian) which use the same ANSI standard character set.

Sino search engine

Sino (which stands for 'size is no object') is an open source, free text search engine which is intended to achieve speed, flexibility, portability and reliability. It exploits the trade-off between disk space and speed, because the size of the concordance (i.e., the index file) built for a set of documents is typically about 40% of the

total size of the documents. This extra overhead for indexing results in fast searching by Sino.

Sino consists of two programs, the indexer 'Sinomake' and the search engine itself. The normal mode of operation of Sinomake is to rebuild the whole concordance. However, it is possible to invoke Sinomake with extra flags to update incrementally the concordance rather than rebuild it, which is much faster than rebuilding the whole concordance.

The Sino search engine program provides several interfaces for developers to interact with it. The most common is the Perl Sino API. Sino also has a flexible search parser which supports various logical connectors in search expressions used in different systems such as Google, Lexis and WestLaw.

LANGUAGES AND LAW IN ASIA

One of the main challenges in the provision of multilingual legal documents is a sheer number of languages in use around the world. From the perspective of this article, the diversity of languages used in Asian legal systems is considerable.

Language	Primary Regions Spoken	Est.
* Chinese (Mandarin)	China, PRC	874
* Hindi-Urdu	India, Pakistan	366
* English	North America, Great Britain, Australia, South Africa	341
Spanish	Latin America, Spain	341
Arabic	North Africa, Middle East	183
* Portuguese	Brazil, Portugal, Angola, Mozambique	176
Russian	Former Soviet Union	167
* Bengali	Bangladesh, India	162
* Japanese	Japan	125
German	Germany, Austria, Switzerland	100
* Korean	Korea	78
French	France, Canada, Belgium, Switzerland, francophone Africa	77
* Chinese (Wu)	China (Shanghai)	77
* Javanese	Indonesia (Java)	75
* Chinese (Yue)	China (Guangdong)	71
* Telugu	South India	69
* Vietnamese	Viet Nam	68
* Marathi	South India	68
* Tamil	South India, Sri Lanka	66
Italian	Italy	62
* Urdu	Pakistan	60

(Est. = Estimated number of native speakers, in millions)
Table adapted from Junker (2003).

Languages spoken and used on the internet

There are more than 6,000 languages still in use today, despite the many that have been lost. The table[5] opposite shows the 20 most popular spoken languages, based on an estimate of the first or primary languages spoken.

Of these 20 languages, 14 are widely spoken in Asia (where asterisks have been added to the table). These include two European languages, English (spoken widely and used in the legal systems of India, Pakistan, Sri Lanka, Bangladesh, Singapore, Brunei, Malaysia and Hong Kong, even though it is not the primary language in some of those speakers), and Portuguese (spoken and used in the legal systems of Macau and Timor Leste). The table does not give the full picture, as it only refers to people's primary (first) language, and therefore (for example) significantly under-estimates the extent to which English is spoken by excluding India from the list of English-speaking countries.

The table demonstrates the diversity of languages spoken around the world, and in Asia, but some popular languages share a common written language. In particular, there are many dialects of Chinese spoken in the People's Republic of China (PRC) as well as by overseas Chinese, three of which are amongst the 20 most popular languages used in the world. There are two common forms of written Chinese that need to be considered, simplified and traditional.

Since this research concerns internet search facilities, it is relevant to also ask which languages are used on the internet. The most recent version of a survey of languages used on the internet (2010)[6] estimates that English (536.6m users, most outside Asia) and Chinese (449.9m users) are by far the two leading languages. Three other languages used significantly in Asia are in the top ten: Japanese (4th with 99.1m users), Portuguese (5th with 82.5m users, most outside Asia), and Korean (10th with 39.4m users). The internet has relatively little penetration in relation to the other most widely used languages in Asia, Hindi and Bahasa Indonesian/Malay (at least according to this survey), but we could expect that to change soon. However, even with the position as it is now, it appears to be of decreasing utility to have an Asia-wide legal information system that only provides information in English, at least from the perspective of the languages spoken by users of the internet. But that is not the only perspective.

Languages used in Asian legal systems, and their representation

Around half of the twenty eight Asian jurisdictions use languages in their legal systems which cannot be

https://doi.org/10.1017/S1472669612000436 Published online by Cambridge University Press

represented in the single byte character sets used for European languages, but require double byte character sets.[7] Some other Asian jurisdictions use single byte character sets to represent the languages in their legal systems, but they are languages which share some of the problems discussed in this article, such as lack of word segmentation,[8] which are also shared with some double-byte languages, and which we are attempting to address.

Partly as a result of these complications, legal texts are often available in various quite different encodings of these national languages. Making the Sino search engine usable for texts from these Asian countries requires exploring approaches to deal with double-byte characters and word segmentation issues, and corresponding development of means of converting other encodings into a standard encoding.

Character sets used in computer systems have undergone significant evolution over the past 50 years. In the 1960s, only unaccented English letters were regarded as important, and they were represented in ASCII, a seven-bit encoding technique which assigns a number to each of the 128 characters used most frequently in American English,[9] and therefore only requires 7 bits for representation (i.e. 128 or 2^7 alternatives). As the need to transfer data between computers increased, this became inadequate. ISO 8859[10] is an eight-bit (or one byte) extension to ASCII developed by ISO (the International Organization for Standardization). It includes the 128 ASCII characters along with an additional 128 characters, such as the British pound symbol and the American cent symbol (i.e. 256 or 2^8 alternatives). Several variations of the ISO 8859 standard exist for different language families, including the various families of European languages, Arabic, Hebrew, and Turkish.

However, ISO 8859 was not sufficient to represent documents from an even wider range of languages which subsequently became available online, especially from Asia. For these languages, the number of characters involved meant that the eight-bit extension was not sufficient. Unicode is an attempt by ISO and the Unicode Consortium to develop a universal character set for electronic text that includes every written script in the world in a consistent manner. Unicode uses 8-bit (single byte), 16-bit (double-byte), or 32-bit characters depending on the specific representation, so Unicode documents often require up to twice as much storage as ISO Latin-1 documents. The first 256 characters of Unicode are identical to ISO Latin-1.

Unicode provides a unique number for every character, no matter what the platform, no matter what the program, no matter what the language.[11] The Unicode Standard defines a fixed-width, 16-bit uniform encoding scheme for written characters and text (Graham, 2000, p 6). The Unicode Version 2 Standard defines 49,194 distinctly coded characters, including characters for the major scripts of the world, as well as technical symbols in common use. The more recent Unicode 5.1.0 contains over 100,000 characters.[12]

English as a linking language in Asian legal systems

Approximately one third of the jurisdictions in Asia use English as an official language in their legal system,[13] sometimes the principal language. In all of these countries this is a legacy of British or US colonialism. Most (perhaps all) of those countries also use other languages for some part of their legal system's operations. India is perhaps the best example of the complexity of the colonial legacy of English (discussed in Greenleaf et al. 2011). India has twenty-two official languages, and somewhere between 150 and 1,500 languages, depending on definitions of language and dialect, (Nilekani, 2008: 77–94). Questions of language always have been and always will be, controversial there. Proposals to adopt Hindi as the only official language in the Constitution met strong resistance from India's southern states, where it was not spoken widely. A Constitutional compromise resulted in Hindi as India's official language, with English to continue in use for all official purposes but only until 1965. However, opposition continued, and in 1967 another compromise was reached providing that 'the use of English as an associate language in addition to Hindi for the official work at the Centre and for communication between the Centre and the non-Hindi states would continue as long as the non-Hindi states wanted it' (Chandra, Mukherjee and Mukherjee, 2008: 123). They conclude that 'English is not only likely to survive in India for all time to come, but it remains and is likely to grow as a language of communication between the intelligensia all over the country, as a library language, and as the second language of the universities' (at 124). It is also likely to retain its privileged, but not exclusive, position in the legal system for some time to come due to various legislative provisions (explained in Greenleaf et al. 2011, under 'Languages other than English'). Other ex-colonies have similarly complex stories.

In addition, in numerous countries throughout Asia where English is not an official language of the legal system, a new development in recent years is that major government-supported efforts are underway to translate a large proportion of each country's important legislation into English and to make the English texts available for free access. This is occurring, or has occurred, often with the assistance of aid-agency funding, in Laos, China, Cambodia, Thailand, Afghanistan, Bhutan, Vietnam and Japan. It is much less common to find significant collections of legal materials within one country in multiple Asian languages, or in European languages other than English. An exception is the use of Portuguese in Macau (where Chinese is also available) and in Timor Leste (where Indonesian or Tetum are also available).

As a result of both the lingering effect of English due to its colonial history and the more recent impetus for English language translations, significant sets of English language legal materials are available from almost all jurisdictions in Asia, as can be demonstrated by an

https://doi.org/10.1017/S1472669612000436 Published online by Cambridge University Press

English language search for almost any legal topic over the AsianLII website. However, English language legal materials are not by themselves fully adequate for legal research in almost all Asian countries. These sets of English language translations are incomplete (except in Hong Kong), rarely constituting more than a modest percentage of the primary legislation. In countries where English is the primary legal language (e.g. India), some legal materials may still only be available in local languages, or only some may be available in translation. It is therefore valuable to have a search engine which can simultaneously search English language texts and can also search material represented in the (often) double-byte representation of the country's language.

Another consequence of the availability of significant quantities of English language legal materials from almost all Asian countries is that it provides ready-made source materials for English to be used as a 'link language', by which translations of concepts from one Asian language to another can be made through the intermediate step of translating concepts from each Asian language into English. This is discussed further in the final section.

DEVELOPMENT OF SINO FOR ASIAN LANGUAGES

For the reasons set out in the previous section, it is desirable for legal information systems in Asian countries to provide a consistent search facility in both English and an Asian language (or more than one), irrespective of which language is the principal language of their legal system. This section and the following, explain progress to date in developing Sino to search Asian languages, particularly languages requiring double-byte representations. We are not aware of multilingual open source search engines being used as yet on free access legal research systems in Asia, other than as described in this paper in relation to Sino.[14]

Purposes of developing Sino for Asian languages

The main purpose of developing Sino to search Asian languages is to assist local organisations in Asian countries to develop free access to law resources in the language of the country using an open source search engine. It will also enable them to build bilingual or multilingual legal research systems if they wish to do so. Developing Sino in this way is probably the most useful contribution that AustLII can make at present to stimulating the development of independently operating LIIs in Asia.

A second reason is that one of the purposes of developing AsianLII is as a comparative law research system across Asia and, eventually, WorldLII as a global comparative law research system. The value of such facilities for the purposes of the APEC, ASEAN, SAARC and other regional groupings, as well as for bilateral trade and investment, is one of the reasons for AsianLII's development (and a justification for its funding). We have developed AsianLII databases of legislation from all Asian countries (except Myanmar), including translations from government sources (but not usually 'official translations') in countries where English is not the principal legal language. This has provided a good start to a comparative legal research system in English across Asia. However, this will be enhanced a great deal if the full texts of legislation are available in the country's language, with links between the versions in both languages. Furthermore, wherever a user is able to do so, the system should provide for searches to be entered in both languages (or as many languages as are known), with a uniform system of search operators and a uniform method of relevance ranking of results.

Unlike in the European Union, it is not realistic in Asia to think about legal research systems with the same materials translated into over 20 languages (as one finds on Eur-LEX). The political and economic conditions of Asia at the present time do not provide the impetus for this, even if they may do in future. It is however increasingly realistic to think about the development of a multi-national and multi-lingual system with the common link between the materials being a version in English. The development of AsianLII, as what we call a multi-bilingual comparative research system, points in that direction.

A general mechanism: Sino u16a approach

We have developed a general mechanism for searching double-byte representations of languages which, in theory, can be utilised with any language, (Asian or otherwise). It may well be that the best long-term answer may be to obtain all content in Unicode and adapt the Sino search engine to search all Unicode representations of languages, but such a solution is some distance away.

The solution described in this paper is more of an interim approach, and is called 'the Sino u16a process'. In simple terms, it can be summarised as follows. A string of text in any language can be converted into a hexadecimal (alpha-numeric or 'flat') representation. The characters 'u16a' (a text string which is rare, almost non-existent, in natural language) are added to each such representation to create a unique string. These u16a 'shadow files' are then used for Sino to search, as a proxy for the original. After the search process is complete, the text in the original language is presented to the user as the content found in the search results. The process is to some extent similar to the process by which PDF image files are made searchable, by use of text files created by optical character recognition (OCR) as the searchable 'shadow files', with the

https://doi.org/10.1017/S1472669612000436 Published online by Cambridge University Press

original image files as the search result presented to the user.

A more detailed explanation is as follows. The core idea behind the u16a representation is that a string in any language using UTF-8 encoding[15] can be converted into an alpha-numeric or 'flat' representation. In other words, a string can be represented as a combination of hexadecimal digits, that is, the range of digits from 0 to 9 and the alphabets from A to F. To ensure that the 'flattened' representation maintains the uniqueness of the string being represented, it is necessary to add another string. From empirical analyses of the concordance of index terms based on the content available from WorldLII, it was discovered that the string 'u16a' was not used as an indexing term (and therefore does not appear in any of those texts) and so was a possible candidate for this task. This was further confirmed by searching general search engines such as Google, which indicated that the string 'u16a' is very rarely used in natural language. Therefore, the characters 'u16a' can be added to any converted alpha-numeric representation to create a unique string. This makes it possible for the term to be retrieved once converted and processed as this new form of representation.

In practice, two files are maintained, the original document and the transformed document with u16a representation. The latter is used as a 'shadow file' for indexing purposes and subsequent searching by Sino as it acts as a proxy for the original. Once indexed, the 'shadow files' become searchable along with other documents supported by Sino. In fact, from Sino's perspective, once transformed, these are 'typical' documents that it can handle for searching. While the u16a encoded documents are used for searching, the text in the original language is presented to the user.

By converting all strings into the u16a representation, the Sino search engine can be used to search documents in other languages outside the ISO range currently supported without having to make any major changes to Sino's core processes.

Example of representing Thai in u16a

The following is an extract of a Thai document and its u16a representation equivalent:

Thai language representation

กรมตรวจบัญชีสหกรณ์ได้มีหนังสือ ที่ กษ ๐๔๐๑/๑๑๐๐๔ ลงวันที่ ๒๘ มิถุนายน๒๕๓๕ ถึงสำนักงานคณะกรรมการกฤษฎีกา สรุปความได้ว่า กลุ่มเกษตรกรทำสวนพงศ์ประศาสน์ (จังหวัดประจวบคีรีขันธ์) ได้จดทะเบียนจัดตั้งกลุ่มเกษตรกรต่อนายทะเบียนกลุ่มเกษตรกรตามประกาศของคณะปฏิวัติ ฉบับที่ ๑๔๑ ลงวันที่ ๑ พฤษภาคม ๒๕๑๕ ต่อมาได้ถูกศาลจังหวัดประจวบคีรีขันธ์

U16A representation:

0E01U16A 0E23U16A 0E21U16A 0E15U16A 0E23U16A 0E27U16A 0E08U16A 0E1AU16A 0E31U16A 0E0DU16A 0E0AU16A 0E35U16A 0E2AU16A 0E2BU16A 0E01U16A 0E23U16A 0E13U16A 0E4CU16A 0E44U16A 0E14U16A 0E49U16A 0E21U16A 0E35U16A 0E2BU16A 0E19U16A 0E31U16A 0E07U16A 0E2AU16A 0E37U16A 0E2DU16A 0E17U16A 0E35U16A 0E48U16A 0E01U16A 0E29U16A 0E50U16A 0E54U16A 0E50U16A 0E51U16A / 0E51U16A 0E51U16A 0E50U16A 0E50U16A 0E54U16A 0E25U16A …

Example searching Thai, Chinese and other languages

The following search query demonstrates how the u16a encoding and the Sino search engine can be used to together in providing searches across a legal information system such as AsianLII or WorldLII that contains documents in a variety of languages. Here, the search for information concerning bankruptcy and insolvency using terms written in English, Thai, Indonesia, Chinese (traditional and simplied), Korean and Vietnamese respectively:

> bankrupt* or insolven* or การล้มละลาย or kepailitan or pailit or 破產 or 破产 or 파산 or Phá sản

Internally, the search is converted to u16a encoding as follows:

> bankrupt* or insolven* or 0e01u16a 0e32u16a 0e23u16a 0e25u16a 0e49u16a 0e21u16a 0e25u16a 0e30u16a 0e25u16a 0e32u16a 0e22u16a or kepailitan or pailit or 7834u16a 7522u16a or 7834u16a 4ea7u16a or d30cu16a c0b0u16a or ph 00e1u16a s 1ea3u16a n

A search over the selected Sino concordances would be conducted based on the u16a representation of the query entered.

The following results page ('By Database') shows the results across a variety of databases available on AsianLII using as the search term (in Thai, Bahasa Indonesia, Chinese and Vietnamese, but omitting English so as to make the results more concise).

How universal is u16a encoding?

In theory, the u16a encoding is of universal application. However, the effectiveness of the Sino u16a process, measured in terms of both database building efficiencies and retrieval speeds, varies between languages. Use of Sino's u16 approach requires an analysis of the structure of each language, and resulting choices in implementation, to obtain the most effective results. Another factor which must be considered is that the resulting storage overhead doubles the storage needed for double-byte languages, and the single-byte languages outside the ASCII range.

We have experimented with making searchable collections of texts in Chinese, Vietnamese, Korean and Thai. Use of the Sino u16a process on representations of

https://doi.org/10.1017/S1472669612000436 Published online by Cambridge University Press

การล้มละลาย or kepailitan or pailit or 破產 or 破产 or Phá sản

1. 安徽高院: 4 documents
2. 北京高院: 6 documents
3. 最高人民法院: 34 documents
4. 青海法院: 2 documents
5. 上訴法庭: 132 documents
6. 終審法院: 14 documents
7. 原訟法庭: 542 documents
8. 區域法院: 86 documents
9. 家事法庭: 9 documents
10. 土地審裁處: 17 documents
11. 憲法文件: 3 documents
12. 香港條例: 1531 documents
13. 香港附屬法例: 1062 documents
14. 法律改革委員會諮詢文件: 12 documents
15. 法律改革委員會報告書: 26 documents
16. 個人資料私隱專員公署行政上訴委員會個案簡述: 1 document
17. 個人資料私隱專員公署投訴個案簡述: 3 documents
18. 實務指示: 10 documents
19. 澳門中級法院: 1 document
20. 澳門終審法院: 6 documents
21. 澳門特別行政區法例: 93 documents
22. 司法院大法官: 7 documents
23. Constitutional Court of Indonesia: 11 documents
24. Keputusan Presiden: 9 documents
25. Peraturan Bank Indonesia: 1 document
26. Peraturan Presiden: 1 document
27. Peraturan Pemerintah: 24 documents
28. Staatsblad: 13 documents
29. Undang Undang: 41 documents
30. Mahkamah Agung Republik Indonesia: 15 documents
31. Pengadilan Tinggi Agama: 3 documents
32. Pengadilan Negeri Jakarta Pusat: 8 documents
33. Thai Legal Materials (in Thai): 100 documents
34. East Timor Regulations promulgated by UNTAET - Bahasa Indonesia: 1 document
35. Laws of Vietnam: 13 documents
36. The Supreme People's Court of Vietnam: Benchbook Online: 56 documents

Chinese language texts results in very fast searching because there are 5,000 or so unique characters (those commonly used) encoded. In contrast, the Thai language is more difficult because its structure is that of 40 alphabetic characters (letters), with no word delimiters, so each letter has to be treated as a separate string with 'u16a' added to it. As a result, Thai u16a concordances are very large, and searches will be relatively slow. However, search efficiency does scale up (at least to collections of 120,000 documents), but not enough testing has yet been done on large data sets to establish whether the result is of practical utility. Other Asian languages such as Japanese, Korean or Hindi will present different problems, and each requires separate testing.

An additional complication is that, to arrive at a u16a representation of a document may involve multiple conversion processes due to the different encoding of the original document. In other words, a document may need to be converted from its source encoding (such as an extended ISO range) into a UTF-8 encoding representation before it can then be converted to u16a for searching on the system. For example, a Chinese document that is originally encoded using the GB 2312 code set will need to be converted to a UTF-8 encoded document in the first instance

https://doi.org/10.1017/S1472669612000436 Published online by Cambridge University Press

(using standard tools such as iconv[16]). A second conversion will then be needed to convert the UTF-8 encoded document into its u16a representation. The availability of convenient and open source conversion software has to be considered for each language.

IMPLEMENTATIONS OF THE u16a SINO APPROACH

As yet, there are two implementations of the Sino u16a process in production, one by the Hong Kong Legal Information Institute (HKLII), and the other on AsianLII. Both use the Chinese language implementation of Sino, to search texts simultaneously in Chinese and in English.

Use by HKLII

The Hong Kong Legal Information Institute (HKLII at http://www.hklii.org) is a free access site for Hong Kong law operated by the University of Hong Kong. The legal system of Hong Kong is bilingual with both English and Chinese as official languages.

Until 2011 HKLII used *mnoGoSearch*, an open source General Public Licence (GPU) search engine designed for the Chinese language. Fung et al. (2011) describe it as follows:

> "It supports Unicode and consists of a built-in dictionary which helps the user to eliminate errors arising from wrong extraction of Chinese words from a document (commonly referred to as 'segmentation errors'). mnoGoSearch supports a wide range of databases. In our case, we had chosen to use MySQL, as it was one of the most reliable databases in the open source community.
>
> mnoGoSearch also consists of an indexer and the search engine itself. The indexer of mnoGoSearch basically extracts sentences delimited by punctuation marks, and extracts strings using its built-in dictionary. All extracted strings are stored as indices in MySQL.
>
> In our experience with mnoGoSearch we had encountered two major problems. Firstly, since its dictionary contained only general Chinese terms, many legal terms contained in the Chinese documents of HKLII were not indexed by mnoGoSearch. Secondly, the searching speed of mnoGoSearch was not satisfactory. Searching simple terms might take up to 10 seconds, and searching more complex Boolean queries might take 30 seconds or more. As a result, we had to constantly fine-tune mnoGoSearch in order to provide an acceptable service to our users. This was done until we experimented with the new version of Sino and found it produced satisfactory results."

This meant that HKLII had to build separate concordances for Chinese (using mnoGoSearch) and English documents (using Sino), as the indexed words in the two languages are all different.

The performance analysis of the u16a implementation of Sino for Chinese conducted by Fung and Pun (Fung et al. 2011) was over databases containing 91,000 Chinese language documents, and 131,000 English language documents.

The results for the building of the concordances were as follows:

	Chinese documents	*English documents*
Total Number of files	91K	131K
Total File Size	1,272MB	1,465MB
Time needed for indexing	2m53s	10m51s
Indexing Speed	441MB/ minute	135MB/ minute
Size of concordance	396MB	862MB
Index ratio	31%	59%

They concluded that Sino indexed the Chinese documents faster than the English documents, and with an index/file size ratio half as large for the Chinese documents. This good result was probably because 'the number of Chinese characters used in legal documents is relatively limited' and 'Chinese characters are repeated more frequently than English words in the documents contained in HKLII'.

Fung and Pun tested whether the u16a representation did provide an accurate method of conversion for searching purposes. They chose at random, 20 names of judges, lawyers and parties to cases, and searched the databases for them. They then checked manually that the names were in fact found in the documents retrieved. In all cases they were found.

In relation to search speed, Fung and Pun tested searches using the 500 Chinese search phrases most frequently used by users of their previous search engine. For exact phrase match searches, the average search time for a Chinese phrase was 0.048 seconds, under optimal conditions (direct interaction with Sino, without network overheads). For 'any of these words' searches (which retrieved many more results) for the same phrases, the average time was 0.103 seconds, with an average of 2,056 documents returned. Searches randomly combining any of these phrases using two of the three connectors 'AND', 'OR' and 'NEAR' (within 50 words) were also tested, and the average search time was 0.097 seconds (OR and NEAR) and 0.096 seconds (AND), with an average number of documents returned being one or zero. They concluded that '[a]ll Boolean

https://doi.org/10.1017/S1472669612000436 Published online by Cambridge University Press

expressions used could be searched within very short time'. They then tested the 'OR' connector to connect multiple phrases (3 or more), and found that even in the most complex case (16 'OR's to connect 17 phrases), the search time was still only 0.830 seconds, with an average of 3,692 documents retrieved, The retrieval time increased in a linear fashion along with the number of 'OR's used, but still less than one second for 16 connectors.

Their overall conclusions about the 'new Sino search engine using the u16a representation' were (Fung et al. 2011):

> "It is fast in both indexing and searching, surpassing the non-western search engines that we had previously encountered. The new Sino search engine has resolved two important problems that HKLII had faced in the past concerning Chinese searching. Firstly, it has avoided the time spent in having to recognise the proper Chinese words contained in the search phrase. As new Sino indexes Chinese documents by character, all search phrases can be handled on the character basis and not on the word basis. Secondly, since u16a representation is alpha-numeric, the new Sino search engine is able to search documents in HKLII in both Chinese and English at the same time".

Use by AsianLII

The second implementation in production is the Chinese language implementation used on AsianLII to provide a comparative law search facility across 74 databases in Chinese from the Peoples Republic (58), Hong Kong SAR (9 from HKLII), Macau SAR (6) and Taiwan (1) at <http://www.asianlii.org/chi/>. These databases can also be searched simultaneously with their corresponding texts in English and Portuguese, from the front page of AsianLII. No equivalent multi-jurisdictional search facility exists elsewhere.

The databases from Macau currently included in AsianLII are in Chinese, Portuguese and (to a very small extent) in English. They can all be searched simultaneously from the Macau home page in AsianLII[17]. It is intended that they will form part of a separate new legal information institute, tentatively named 'MacauLITES', which will be operated by the University of Macau. As with HKLII this proposed LII will utilise the capacity of Sino's u16a representation.

UNRESOLVED ISSUES AND FUTURE WORK

Our work on the issues of searching law in multiple Asian languages is only at an early stage. In relation to traditional Chinese, used concurrently with English data, the Sino u16a approach has produced an effective solution which has been independently tested, put into production, and works well. For other Asian languages, full demonstrations of the u16a approach are still to be completed, but the approach seems promising, and there is a need for a multi lingual search engine. To conclude, we outline one key problem which needs to be addressed with many Asian languages, word segmentation, and then considers the uses that can be made of the u16a approach for cross-lingual searching.

A key problem: Word segmentation

In Western languages such as English, words are generally explicitly delimited by white spaces. Some equivalent to 'words', significant groupings of characters carrying meaning, are needed for efficient text processing tasks such as searching and information retrieval. However, non-Western languages and in particular many Asian languages such as Chinese, Japanese, Korean and Thai do not exhibit this linguistic feature generally referred to as the 'word segmentation' problem. Being able to accurately identify the word boundary is a core component of addressing this issue. The 'word segmentation' problem has been extensively investigated, for example by Peng et al. (2002); Pun, Chong and Chan (2003), and Nguyen et al. (2006).

As an example, in the case of Chinese, which is based on ideographic writing, its system does not use space or any other delimiter as word boundaries. Pun, Chong and Chan (2003) provides the following example:

我們要發展中國家用電器

One way to segment this sentence is:

我們	要	發展	中國	家用電器
We	want	to develop	China's	home electrical appliances

https://doi.org/10.1017/S1472669612000436 Published online by Cambridge University Press

Another way of segmenting this sentence which has a different meaning is:

我們	要	發展中國家	用	電器
We	want	developing countries	to use	electrical appliances

There are many approaches to the word segmentation problem. These can be basically divided into character-based and word-based approaches (see, for example, Foo & Li 2004; Pun, Chong and Chan, 2003; Haruechaiyasak, Kongyoung and Dailey, 2008). In character-based approaches, the focus is on extracting a certain number of characters as the basis for segmentation. Character-based approaches can further be classified into single-based (uni-gram) or multi-based (n-gram) approaches (Nguyen et al. 2006).

Word-based approaches can be subdivided into dictionary-based; and statistics-based or machine-learning based. The dictionary-based approach relies on dictionaries that contain the most common words and employs heuristic rules to recognise compound words not found in dictionaries. Using this approach, the system's performance in segmentation depends greatly on the comprehensiveness of the dictionary (Pun, Chong and Chan, 2003). The previous search engine used by HKLII (mnoGoSearch) essentially adopts a dictionary-based approach. The statistics-based approach relies on statistical information such as term, word and character frequencies to create a table of words and their corresponding weights. These weights are used to compute the score for a potential segmentation of a sentence (Nguyen et al. 2006). If a sentence can be segmented in more than one way, the segmentation with the highest score, computed based on the weights of the words identified therein, will be selected (Pun, Chong and Chan, 2003). This means that the effectiveness is dependent upon a particular training set.

The Sino u16a representation does not deal with this issue directly. It can be considered to adopt a character-based uni-gram approach to the segmentation issue, which by its nature will produce some ambiguous results. While there are unresolved issues in relation to performance and word segmentation, requiring further research, our pragmatic starting point is that to have a working multilingual comparative search system is more significant that to solve all theoretical issues. However, this requires us to demonstrate that, as with the HKLII example, the u16a approach can produce results of significant practical utility with other languages.

Cross-lingual searching with English as a link language

The availability of documents in multiple languages within the one search system makes it highly desirable to be able to provide 'cross-language' searching. Cross-language or cross-lingual searching or information retrieval (CLIR) can be considered as the retrieval of documents in a language other than the language of the request or query.

In discussing cross-lingual searching ('CLIR' or 'cross-lingual information retrieval'), three types of search tasks have been identified (Kishida et al, 2004): SLIR (single language IR); BLIR (bilingual CLIR); and MLIR (multilingual CLIR), as follows:

(i) SLIR is where the language of the search topics (usually determined by the linguistic capacity of the human searcher) is identical to that of the documents (i.e. this is not a cross-lingual task).
(ii) BLIR denotes that a document set in a single language is searched using topics in a different language (for example, using English topics to search Chinese documents).
(iii) MLIR denotes a search task where the target collection consists of documents in two or more languages (for example, searching a multilingual collection for Chinese topics).

Ideally, the use of Sino to search AsianLII (or HKLII or MacauLITES or other implementations) should involve effective MLIR. One possible approach to MLIR is to use bilingual dictionaries available from a number of countries in order to make 'inferences' concerning mapping of search terms between languages. These 'mappings' can then be used to facilitate query translation. For example, bilingual legal dictionaries using English as one of the languages already exist in Hong Kong[18] and Japan[19] (Matsuura, 2012; Toyama and Yasuhiro, 2012; Sekine 2012) and are available for free-access. One is also available from Korea[20] (Hong, 2012), but it is not clear whether it is available for free access uses by other research teams. A Japanese-English-Chinese dictionary relevant to law in Taiwan is also under development[21] (Hwang and Shee, 2012), as part of a joint Taiwan-Japan-Korea-PRC research project.[22] We expect that use of such dictionaries, preferably in collaboration with the research teams developing them, will provide AustLII and AsianLII with an efficient way to develop a basic 'cross-lingual' facility within a comparative legal research facility. AustLII's aim is only to develop automated (or perhaps semi-automated) translation of search queries, not to do more ambitious translations.

However, we have no expectations that translations of search terms is a simple matter. There are a number of limitations including difficulties due to different legal traditions, such as the differences between civil law and common law jurisdictions. Also, it may be difficult to have translation from one language and mapping directly across to another as they are likely to be presented with a number of possible translations; the issue of *ambiguity* (Zhou et al. 2008). The dictionary mapping and inferencing approach is only likely to work for core legal concepts and will not be able to handle the subtleties of different legal frameworks and terms. This problem is often referred to as the *coverage* problem or more specifically, 'out-of-vocabulary' (OOV) problem. For

https://doi.org/10.1017/S1472669612000436 Published online by Cambridge University Press

example, Zhou et al. (2008) discuss these issues in relation to English-Chinese cross-language retrieval. Other approaches use statistical techniques and more sophisticated translation frameworks (Liu, Jin & Chai, 2006; Gao, Nie & Zhou, 2006).

From AsianLII's perspective, we expect that the initial approach will be the relatively simple one of using a synonym list for common legal terms as the first step for cross-language searching. Using the Sino search engine, a sino_synonym file can be created with comma (or space) separated entries denoting translations of the term in different languages. For example, in relation to European languages, the EuroVoc Thesaurus <http://eurovoc.europa.eu/> could be used to develop the following synonyms lists for six European languages for Sino to implement:

election, elezione, wahl, valg, elecciones, vaalit

constitution, costituzione, verfassung, perustuslaki

These can be further supplemented by a Chinese-English bilingual dictionary to produce:

election, elezione, wahl, valg, elecciones, vaalit, 選擇,選挙

constitution, costituzione, verfassung, perustuslaki, 憲法, 憲法

A search using any one of these terms would then find any documents using any of the other terms in the synonym list. Whether this approach is good enough to be of practical utility, and in particular whether it can be used to generate useful compound searches (at least AND and OR connectors) remains to be seen.

AustLII's initial testing of this approach will take place in the second half of 2012 in joint research on with the Hong Kong Legal Information Institute (HKLII), using the Hong Kong government Chinese-English legal dictionary to generate MLIR searches over HKLII and AsianLII.

Lack of one-to-one correspondence in versions of Chinese

In relation to Chinese, one unresolved issue noted by Fung and Pun was that because HKLII contains only documents in traditional Chinese and English, but not in simplified Chinese, searches in simplified Chinese were therefore not tested by them (Fung et al. 2011: [4]). If a user attempted to search HKLII using a search phrase in simplified Chinese, it would first have to be converted to traditional Chinese, and then to the u16a representation. The problem is that the mapping of simplified Chinese to traditional Chinese is a one-to-many mapping, and therefore difficult to automate if relevant results are not to be missed. Conversely, if databases in simplified Chinese are developed, then if searches are attempted in traditional Chinese, they must first be converted to simplified Chinese, which involves a many-to-one mapping, with the risk that irrelevant search results will be returned. This scenario can be considered as a particular instance of the MLIR issues discussed above. Some of the specific complexities of Chinese to Chinese conversion are discussed in Halpern and Kerman (1999). Conceptual mappings via the use of bilingual dictionaries discussed previously, offer a partial solution to this issue (Halpern and Kerman, 1999).

Expanding AsianLII's Asian languages

In order to allow testing of the Sino u16a representation, AustLII will now add significant collections of legal texts in Japanese and Korean, and expand the existing collection in Thai and Vietnamese. If tests of the u16a versions of these texts show effective search results, then it may be possible to use other available bilingual legal dictionaries to extend the generation of automated search synonyms to test more MLIR covering multiple Asian languages.

Footnotes

[1] AustLII website <http://www.austlii.edu.au/techlib/software/sino/>

[2] Other LIIs using Sino include AsianLII, BAILII, CommonLII, CyLaw, HKLII, LiberLII, LII of India, NZLII, PacLII, SafLII, ULII and WorldLII The full names of these LIIs, and links to their sites can be found from the website of the Free Access to Law Movement (FALM) <http://www.fatlm.org/>

[3] Hong Kong SAR and Macau SAR have legal systems largely separate from that of the PRC.

[4] Wikipedia 'List of Search Engines' page, subheading 'Open source search engines', at <http://en.wikipedia.org/wiki/List_of_search_engines#Open_source_search_engines>

[5] This table is modified from Yunker, J (2003) *Beyond Borders: Web Globalization Strategies*, New Riders, p32. This is sourced from *Ethnologues*, 14th Edition, 2000 <http://www.ethnologue.com/>. The table has been modified by the addition of the asterisks and the rewording 'francophone Africa'.

[6] See 'Top Ten Languages in the Internet' in World Internet Statistics at <http://www.internetworldstats.com/stats7.htm>, June 2010

[7] Including Japan, People's Republic of China, Hong Kong SAR, Macau SAR, North Korea, South Korea, Taiwan, India, Nepal, Bhutan, Myanmar (Burma), Sri Lanka, Maldives, Bangladesh

[8] Thailand, Lao PDR, Cambodia, Pakistan, Afghanistan, Vietnam

https://doi.org/10.1017/S1472669612000436 Published online by Cambridge University Press

[9] The remaining bit in a one byte representation was the cause of many subsequent problems, because there was no standard as to how it should be used (prior to Unicode), so many representation of languages that could not be accommodated by the 7 'ASCII bits' (code spaces 0–127) made inconsistent use of the 8th bit (code spaces 128–255), making transfer of data between computers often impossible.

[10] ISO 8859, *Information processing – 8-bit single-byte coded graphic character sets*

[11] <http://www.unicode.org/standard/WhatIsUnicode.html> (as at 25 April 2004).

[12] <http://www.unicode.org/versions/Unicode5.1.0/> (as at 10 February 2009); now updated to version 6.1.0 on 31 May 2012

[13] Including Pakistan, India, Sri Lanka, Bangladesh, Malaysia, Singapore, Brunei, and Papua-New Guinea and Hong Kong

[14] A survey of open source search engines has not been done. However, *mnoGoSearch* used by HKLII, discussed later, does have the capacity to search Chinese, Thai and some other Asian languages.

[15] See, for example, Lunde (2009, p 206) for an explanation of the UTF-8 encoding form.

[16] See <http://www.gnu.org/software/libiconv/> (as at 1 June 2012)

[17] See <http://www.asianlii.org/resources/2499.html> for the 16 databases, six of which are in Chinese.

[18] See <http://www.legislation.gov.hk/eng/glossary/homeglos.htm> (as at 1 June 2012) for an English to Chinese dictionary of legal terms relevant to Hong Kong law, developed by the Hong Kong government.

[19] See <http://www.japaneselawtranslation.go.jp/dict/download?re=02> (as at 1 June 2012) for an English to Japanese dictionary of legal terms relevant to Japan, developed by the research team at Nagoya University led by Prof Y Matsuura, who leads the Japan Legal Information Insitutute (JaLII - http://jalii.law.nagoya-u.ac.jp/enindex).

[20] This dictionary has been developed by the Korean government. A Korean-Chinese legal dictionary is also under development (Hong, 2012).

[21] Developed by Taiwan Legal Information Institute (TaiwanLII – http://www.taiwanlii.ccu.edu.tw/) led by Prof H-L Shee.

[22] Cross-Territory Research Consortium on Legal Information & Comparative Studies

References

Chandra, B, Mukherjee, M and Mukherjee, A. (2008) *India Since Independence*, Penguin, 2008

Fung, A, Pun, K, Chung, P and Mowbray, A. (2011) 'Searching in Chinese: The Experience of HKLII', paper presented at *Law via the Internet Conference*, Hong Kong, June 2011, available at <http://www.hklii.hk/conference/paper/1C3.pdf>

Foo, S and Li, H. (2004) 'Chinese word segmentation and its effect on information retrieval', *Information Processing and Management* 40(1), 161–190

Gao, J, Nie, J and Zhou, M. (2006) 'Statistical query translation models for cross-language information retrieval', *ACM Transactions on Asian Language Information Processing* 5(4) 323–359

Graham, T. (2000) *Unicode: A Primer*, M&T Books, 2000

Greenleaf, G. (2011) 'Free access to legal information, LIIs, and the Free Access to Law Movement', Chapter in Danner, R and Winterton, J (eds.) *IALL International Handbook of Legal Information Management*. Aldershot, Burlington VT, 2011, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1960867

Greenleaf, G, Chung, P and Mowbray, A. (2007) 'Challenges in improving access to Asian laws: the Asian Legal Information Institute (AsianLII)' (2007). *Australian Journal of Asian Law*, 9(1).

Greenleaf, G, Mowbray, A and Chung, P. (2011) 'AustLII: Thinking locally, acting globally' (2011) *Australian Law Librarian*, pgs 101–11, available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1960878

Greenleaf, G, Mowbray, A and Chung, P. (2010) 'AustLII in 2010 – A snapshot at age 15' AustLII, December 2010, available at: http://www.austlii.edu.au/austlii/publications/2010/1.pdf

Greenleaf, G, Vivekanandan, VC, Chung, P, Singh, R and Mowbray, A. (2011) 'Challenges for free access in a multi-jurisdictional developing country: building the Legal Information Institute of India' *SCRIPTed* 8(3), 2011, University of Edinburgh School of Law; also available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1975760.

Halpern, J and Kerman, J. (1999) 'The pitfalls and complexities of Chinese to Chinese translation' in *Proceedings of Machine Translation Summit VII*, 13–17 September 1999, Singapore. http://www.mt-archive.info/MTS-1999-Halpern.pdf

Haruechaiyasak, C, Kongyoung, S and Dailey, M. (2008) 'A comparative study on Thai word segmentation approaches', *The fifth International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, 125–128.

Hong, S-J. (2012) 'Development of the online legal information system and the e-Legislation system in Korea', p. 158–62 in Shee (Ed), 2012

Hwang, R-H and Shee, H-L. (2012) 'Making legal information smart, friendly and inspiring', p. 163–201, in Shee, H-L (Ed), 2012

Kishida, K, Chen, K, Lee, S, Chen, H, Kando, N, Kuriyama, K, Myaeng, SH and Eguchi, K. (2004) 'Cross-lingual information retrieval (CLIR) task at the NTCIR workshop 3', *SIGIR Forum* 38(1) (Jul. 2004), 17–20. DOI: http://doi.acm.org/10.1145/986278.986281

https://doi.org/10.1017/S1472669612000436 Published online by Cambridge University Press

Liu, Y, Jin, R and Chai, J. (2006) 'A statistical framework for query translation disambiguation', *ACM Transactions on Asian Language Information Processing*, 5(4), 360–387

Lunde, K. (2009) *CJKV Information Processing*, 2nd ed, Oreilly & Associates

Matsuura, Y. (2012) 'The principle of "mutual exploitation" and collaborative approach to comparative study of laws in East Asia', p. 58–71, in Shee, H-L, 2012

Nilekani, N. (2008) *Imagining India: Ideas for a New Century*, Allen Lane (Penguin Press), 2008, p. 77

Nguyen, TV, Tran, HK, Nguyen, TTT and Nguyen, H. (2006) 'Word Segmentation for Vietnamese Text Categorization: An online corpus approach' in *Proceedings of 4th IEEE International Conference on Computer Science – Research, Innovation and Vision of the Future 2006 (RIVF'06)*, p. 172–178

Peng, F, Huang, X, Schuurmans, D and Cercone, N. (2002) 'Investigating the relationship between word segmentation performance and retrieval performance in Chinese IR' In *Proceedings of the 19th international Conference on Computational Linguistics – Volume 1* (Taipei, Taiwan, August 24 – September 01, 2002). International Conference On Computational Linguistics. Association for Computational Linguistics, Morristown, NJ, 1–7. DOI: http://dx.doi.org/10.3115/1072228.1072376

Poulin, D, Mowbray, A and Lemyre, P-P. (2007) 'Free Access to Law and Open Source Software' *Handbook of Research on Open Source Software* St. Amant & Still (Eds) Information Science Reference, Hershey – New York 2007, available at: http://www.informationjuridique.org/docs/publications/freeaccess2007.pdf

Pun, K. (2003) 'Processing Legal Documents in the Chinese-Speaking World: the Experience of HKLII' *Proc. Law via Internet Conference*, 2003

Pun, K. (2004) 'Cross-Referencing for Bilingual Electronic Legal Documents in HKLII', *Proc. Law via Internet Conference*, 2003

Pun, KH, Chong, CF and Chan, Vivien. (2003) 'Processing Legal Documents in the Chinese-Speaking World', [2003] CompLRes 4, *Proc. Law via Internet Conference*. http://www.austlii.edu.au/au/other/CompLRes/2003/4.html

Sekine, Y. (2012) 'The Development of a Translation Memory Database System for Law Translation', *Invitation to Translation Studies* 7, 79–88.

Shee, H-L. (Ed). (2012) *Proc. International Conference on Legal Information and East Asian Law: Theories, Practices and Prototypes*, National Chung Cheng University Department of Law, Taiwan, June 2012

Toyama, K and Sekine, Y. (2012) 'Development of translation memory database system for law translation', p. 153–157, in Shee, H-L (Ed), 2012

Zhou, D, Truran, M, Brailsford, T and Ashman, H. (2008) 'A Hybrid Technique for English-Chinese Cross Language Information Retrieval', *ACM Transactions on Asian Language Information Processing (TALIP)* 7(2) (June 2008), 1–35. DOI: http://doi.acm.org/10.1145/1362782.1362784

Biographies

Philip Chung is a Senior Lecturer in Law, University of New South Wales (UNSW) and Executive Director of AustLII, where he has held senior positions since 1996. He has degrees in law and computer science (operations research), manages AustLII's databases and is the co-developer of the u16a extension to Sino.

Andrew Mowbray is Professor of Law and Information Technology, University of Technology, Sydney (UTS) and a co-founder and Co-Director of AustLII since 1995. He has degrees in law and computer science, and is the author of AustLII's open source Sino search engine, co-developer of its u16a extension, and author of other AustLII software such as the suite of LawCite tools for citation mining and citator creation.

Graham Greenleaf AM is Professor of Law & Information Systems, University of New South Wales (UNSW) and a co-founder and Co-Director of AustLII since 1995. He has qualifications in law, and is a Fellow of the Australian Computer Society. In 2010 he was made a Member of the Order of Australia for his contributions to free access to legal information and to privacy protection.

https://doi.org/10.1017/S1472669612000436 Published online by Cambridge University Press