From the cultivation of comparable corpora to harvesting from them: A quantitative and qualitative exploration

Benjamin K. Tsou
City University of Hong Kong
Hong Kong University of Science and Technology
Chilin (HK) Ltd
btsou99@gmail.com

Kapo Chow
Chilin (HK) Ltd
kapo.rclis@gmail.com

Abstract

This paper reports on a relatively new but important area involving: (1) The corpus cultivation of comparable multilingual patents for, (2) The building of a large-scale parallel sentence corpus, and thence, (3) The harvesting from them of useful language resources for NLP and other applications. Three major efforts are reported: (a) The sourcing and cultivation of a large corpus of bilingual and comparable patents suitable for diverse applications in strategic NLP, (b) The mining of high-quality bilingual aligned sentences from the comparable patents which contain many technical terms, and (c) The extraction of bilingual multi-word expressions (MWEs) from the parallel sentences in (b), where there is often no simple isomorphic cross-lingual correspondences among the lexical items. An analysis is provided on how over 1 million very high quality lexical entries consisting of bilingual multi-word expressions and their multiple renditions have been harvested in the initial and expanding version of a derived MWE database. This has followed a series of rigorous winnowing of an initial large database of 10 years of English and Chinese patents consisting of more than 5,000 million English words and 12,000 million Chinese characters. Some issues on efficacy in data curation to maximize both quantity and quality are raised, as well as an outline of how the MWEs with their multiple renditions are put to good use by translators and trainers of translators.

1 Introduction

To be useful for NLP, the size of corpora must be quite large. In recent decades, there have been interests in working with big and related corpora in different languages, where the degree of comparability may range widely from parallel texts to even unrelated texts (Sharoff et al., 2013). Such interests were boosted by, for example, the official production of bilingual corpora (Germann, 2001; Koehn, 2005). Common examples are voluminous bilingual parliamentary records and parallel bilingual legal codes, for example, from governments or international organizations, as well as the availability of multilingual comparable corpora from Wikipedia. The availability of other bilingual texts which are comparable, if not parallel, has boosted considerably advances in machine translation and other NLP efforts (Sharoff et al., 2013; Zhao and Vogel, 2002; Resnik and Smith, 2003; Munteanu and Marcu, 2005; Wu and Fung, 2005; Smith et al., 2010). However, comparable corpora as well as similar and useful monolingual corpora also exist in other relatively untapped domains and areas which are of strategic importance. This is especially the case with patents in the cross-lingual context which impinge on complex global trade and economic competition as their content involves advanced scientific and technological developments which require comprehensive but succinct description and where ownership of intellectual property rights may be contentious and have to be protected legally.

2 Corpus Cultivation: From Quantity to Quality

2.1 Stage A: From Big Data to Useful Data

There can be several stages in the long work flow to go from Big Data to useful data and, it is no simple task which can be accomplished semi-automatically. It is often assumed that the more data the better, and that the Age of Big Data would provide easy and theoretically limitless access to data. In the case of patents, they are documents which usually contain much useful information and new technical terms, and may be available in more than one language because inventions described in a patent may only be best protected in the country where it is filed. As a result, an applicant who wishes to protect his invention would file the patent in other countries in the local languages. Usually, the two or more versions should be parallel but there could be also textual variations to tactfully protect the legal rights of the technical content and because of non-uniformity in human efforts. There are also cross-references to other relevant patents, and quite often there are bilingually paired patents or bilingually paired textual segments within clusters of comparable patents.

At the start, we took 10 years of Chinese and English patents officially published around the turn of the millennium. The combined size of these English and Chinese patents is very large:

English: 5,351.7M words
Chinese: 12,001M characters

This enormous quantity is based on the number of English and Chinese patents published during this decade: English patents: 840,027 documents and Chinese patents: 967,686 documents, and based on the average size of more than 300,000 English and Chinese patents which have been analyzed. (Lu et al.,2016)

2.2 Stage B: Corpus Cultivation of Comparable patents

The identification and winnowing of comparable patents from stage A begins with the meta-information of the patents. Essentially the cross-references in the section “Worldwide applications” are examined.

From the official 2009 website of the State Intellectual Property Office (SIPO) in China, about 200K Chinese patents were found to have links with previously filed PCT applications in English and we crawled their bibliographical data, titles, abstracts and the major claim from the Web. Other claims and descriptions were also added in the processing. All PCT patent applications are filed through WIPO. Drawing on the Chinese patents mentioned above, the corresponding English patents were searched from the WIPO website to obtain relevant sections of the English PCT applications, including bibliographical data, title, abstract, claims and description. About 80% (160K) of the Chinese patents have corresponding English ones and a total of about 340K comparable bilingual patents form the initial base corpus of comparable patents. (Lu et al., 2015, 2016)

These cultivation efforts involved considerable manual efforts and have yielded the following combined size of textual content .

English: 1,020.4M words
Chinese: 1,986.4M characters

2.3 Stage C*1*: Sentence Alignment 1

Following stage B our preliminary efforts produced a drastically reduced set of bilingual English-Chinese parallel sentences by means of iterative bilingual sentence alignment.

First we use a bilingual seed dictionary to preliminarily align the sentences in each section (abstract, claims, descriptions) of the comparable patents, and perform filtering using length-based (Gale and Church, 1991) and dictionary-based scores. The dictionary-based similarity score P_d of a sentence pair is computed based on a bilingual dictionary as follows (Utiyama and Isahara, 2003):

$$p_d(S_c,S_e)=\frac{\sum_{w_c\in S_c}\sum_{w_e\in S_e}\frac{\gamma(w_c,w_e)}{\deg(w_c)\deg(w_e)}}{(l_e+l_c)/2}$$

where w_c and w_e are respectively the word types in Chinese sentence S_c and English sentence S_e; l_c and l_e respectively denote the lengths of S_c and S_e in terms of the number of words; and $\gamma(w_c,w_e)$ = 1 if w_c and w_e is a translation pair in the bilingual dictionary or are the same string, otherwise 0; and

$$\deg(w_c)=\sum_{w_e\in S_e}\gamma(w_c,w_e)$$

$$\deg(w_e)=\sum_{w_e\in S_c}\gamma(w_c,w_e)$$

For the bilingual dictionary, we drew from three publications: viz, LDC_CE_DIC2.0 (http://projects.ldc.upenn.edu/Chinese/LDC_ch.htm), bilingual terms in HowNet (http://dict.cnki.net/) and the bilingual lexicon in Champollion (Ma, 2006). We then removed sentence pairs using length filtering and ratio filtering by two means: 1) For length filtering: if a sentence pair had more than n words in the English sentence or more than m characters in the Chinese one, it was removed; 2) For length ratio filtering: we discarded sentence pairs with

Chinese-English length ratio outside the range of 0.8 to 1.8. The parameters here were set empirically. We further filtered the parallel sentence candidates by learning an IBM Model-1 algorithm (Brown et al., 1993; Och and Ney, 2004) on the remaining aligned sentences and computing the translation similarity score P_t of sentence pairs by combining the translation probability value of both directions (i.e. Chinese->English and English->Chinese) based on the trained IBM-1 model (Moore, 2002; Chen, 2003; Lu et al, 2009). It is computed as follows:

$$p_t(S_c, S_e) = \frac{log(P(S_e \mid S_c)) + log(P(S_c \mid S_e))}{l_c + l_e}$$

where $P(S_e \mid S_c)$ denotes the probability that a translator will produce S_e in English when presented with S_c in Chinese, and vice versa for $P(S_c \mid S_e)$. Sentence pairs with similarity score P_t lower than a predefined threshold are filtered out as incorrect aligned sentences.

After the end of stage C1, we have an initial bilingual sentences corpus with the following combined size.

English: 355.3M words
Chinese: 628.2M characters

2.4 Stage C2: Sentence Alignment 2 (with enriched bilingual glossary)

Cyclical sentence alignment is applied with the help of an enriched bilingual glossary. This has enhanced the size of bilingual sentence corpus as follows.

English: 989.2M words
Chinese: 1,914.6M characters

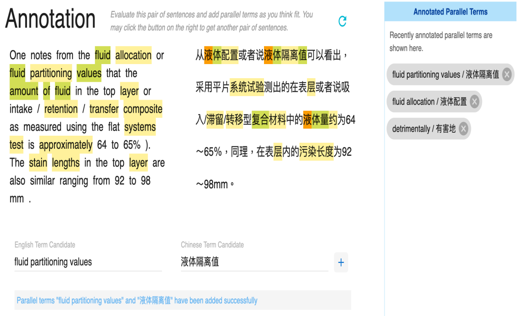
It is noteworthy that the results show near maximal recall as can be seen from the ceiling effect when compared with the results of stage B. At the same time, it reflects on the efficacy of the sentence alignment efforts in stage B.

2.5 Semi-automatic Curation Efforts

Whereas stage A identified patents whose contents are parallel, if not comparable, stage B identified linguistic segments within them which are comparable on objective and statistical basis.

From stage B, automatic multi-word expression extraction produced high-frequency bilingual terms. But within the parallel corpus there were residual and low frequency but valid useful terms for stylistic or other reasons. They could be further exploited. Because of their rarity and sparsity, there was no effective automatic means to extract these low-frequency bilingual terms. We thus developed a computer-assisted bilingual term extraction system to allow annotators to perform this task in a simple but effective manner. The system has been a web-based online system for annotators to mine unextracted term pairs in the following steps:

First, pairs of aligned sentences are randomly selected from the pool of about 40M candidate sentence pairs. For these pairs, all recognized bilingual terms from the current bilingual glossary are highlighted. To enhance visual presentation, words that are covered by multiple bilingual terms are highlighted in a darker shade. This allows annotators to focus on unextracted terms at a glance, and highlight them in both languages, thereby adding the selected term pair to the database. Once annotation of all unextracted bilingual terms are completed, the annotator can choose to view another random pair of sentences from the system, and repeat the same examination and annotation process.



Annotation

Evaluate this pair of sentences and add parallel terms as you think fit. You may click the button on the right to get another pair of sentences.

One notes from the fluid allocation or fluid partitioning values that the amount of fluid in the top layer or intake / retention / transfer composite as measured using the flat systems test is approximately 64 to 65%). The stain lengths in the top layer are also similar ranging from 92 to 98 mm .

从液体配置或者说液体隔离值可以看出，采用平片系统试验测出的在表层或者说吸入/滞留/转移型复合材料中的液体量约为64~65%，同理，在表层内的污染长度为92~98mm。

English Term Candidate: fluid partitioning values

Chinese Term Candidate: 液体隔离值

Parallel terms "fluid partitioning values" and "液体隔离值" have been added successfully

Annotated Parallel Terms

Recently annotated parallel terms are shown here.

fluid partitioning values / 液体隔离值

fluid allocation / 液体配置

detrimentally / 有害地

From the sample user interface shown above, it can be seen that two pairs of terms, “fluid allocation” / “液体配置” and “fluid partitioning values” / “液体隔离值”, are newly marked and added to the database. The system underlying this curation process is outlined in the next section, which focuses on the systematic extraction of linguistically well-formed multi-word expressions.

2.6 Stage Dn: Multi-word Expressions

We implemented an automatic procedure to acquire the bilingual phrases from the above parallel sentence pair corpus on the basis of Tian et al. (2011, 2014) and discussed in Tsou et al. (2017b). We first derived the phrases from monolingual data, i.e. the source sentences of parallel data, by considering the possible n-gram of words. For the high frequency words or phrases, we evaluated how likely a phrase could be constructed by two (words or) phrases p_i and p_j, by considering the following significant score:

$$\text{Score}(P_1, P_2) = \frac{f(P_1 \oplus P_2) - \mu_0(P_1 - P_2)}{\sqrt{f(P_1 \oplus P_2)}}$$

where $f(\cdot)$ and $\mu_0(\cdot)$ were the frequency and the mean under null hypothesis of independence of two phrases (El-Kishky et al., 2014). $\oplus$ was the concatenation operator. The equation computes the number of standard deviations away from the expected number of occurrences, and this score could be considered a generalization of the t-statistic for identifying dependent bigrams. To extend the identified phrases to its bilingual, we first derived the word alignment information for the bilingual data using the model proposed by Dyer et al. (2013). The word alignments acted as vital information and were used to project the phrase boundaries from the source sentences to the target side of the bilingual data (Zeng et al., 2014). For those unaligned words or phrases, we simply ignored them from the induction process. This bilingual phrases extraction model was based on an unsupervised approach, where all the statistics were automatically derived from a given parallel corpus aligned at sentence level.

The bilingual multi-word expressions obtained were fed into an online translation engine for further automatic evaluation. Each monolingual part of a pair of bilingual terms was input to a translation engine whose output was compared with the other part of the bilingual pair in terms of Levenshtein distance (LD) (Haldar et al. 2011). The results were as follow: LD=0:11.9%, LD=1:25.5%; LD=2:29.3%; LD=3:11.9%; LD=4: 11.0%; LD≥ 5:10.3%.

We noted that from the results, only 11.9% were identical with online translation results. The remaining nearly 90% were not identical but were still potentially valid entries. They were more valuable because they reflected the actual alternate language use in this particular domain which would be usually ignored by automatic means and also was not readily available through any public resources. Further empirical studies showed that lower edit-distance entries require less manual modifications. When the edit distance was 1 or 2, about 65% were valid entries without modifications and 5% more could be useful after manual modifications. For distance 3 or 4, about 55% were valid entries, while about 10% could become useful after modifications.

Our efforts so far have produced over 6 million MWE candidate entries. Furthermore, mostly straightforward manual checks have already yielded 1 million good bilingual MWEs and more items are expected after iterative processing. The human efforts have mostly involved the pruning of redundant constituents and in some cases the recovery of missing constituents.

We note there is a noticeable drastic reduction in going from the parallel aligned sentences and sentence fragments to bilingual terms, including MWE's. Following further filtering and human supervision are found in the following sub-corpus only linguistically well-formed expressions.

English: 2.95M words
Chinese: 5.89M characters

2.7 Pairing Bilingual Terms

Based on the bilingual MWE database, we have constructed a cross-lingual search system – *Chilin PatentLex*. The following are some examples of search results. Based on the meta information of each patent, we are able to provide insightful statistics through the search query, as can be seen in Table1.

2.7.1 "Channel" – Alternate Renditions in Chinese Patents

PatentLex (%)	PatentLex (%)	PatentLex (%)
道 (48.58)	沟槽 (0.58)	水道 (0.01)
信道 (30.89)	管道 (0.37)	槽钢 (0.01)
通道 (10.92)	道宽 (0.12)	沟渠 (0)
频道 (3.06)	渠道 (0.06)	通道化 (0)
槽 (2.55)	波道 (0.02)	海峡 (0)
沟道 (1.98)	途径 (0.02)	管箱 (0)
腔 (0.71)	路线 (0.01)	通风槽 (0)
		窜槽 (0)

Table 1 Alternate Chinese Renditions for "*Channel*"

In Table 1, we note that the English term "channel" has 22 renditions in Chinese and that the percentage distribution of actual usage of each alternate rendition shows considerable variations. Among the 22 actual alternate renditions, 3 are used in 10% or more of the total, while the majority has frequency less than 1% (0% indicates very low usage of less than 0.01). While this is exhaustive for the 10 years of Chinese-English patent we collected, it may be noted that the different IPC domains have not been equally represented in patents. Given further cultivation of the comparable database and also breakdown according to IPC domains, we expect the distribution of the alternate renditions to change and be more balanced.

From the current database, we can see that the highest percentage of usage comes from the single-character word 1."道", followed by 2. "信道" and 3."通道". It is worth noting that since the percentage comes from figures gathered from string matching of the rendition with the parallel sentence pairs, the rendition “信道” will also contribute to the counting of“道”. This may lower the *precision* but can increase *recall* and lead to more relevant authentic example sentences for users to examine.

We can further retrieve all the example sentence pairs containing MWE’s of “channel” with various renditions from different IPC domains (see examples in Table 4), and even with low frequencies of usage, so that the needs of the translator may be met.

Furthermore, beyond the renditions of the search keyword “channel”, the search engine will also return other fuzzy results with MWEs containing “channel”, again with their respective renditions and distribution, as can be seen in Table 1 and Table 2.

English Matched Term	Chinese Renditions (%)
a corresponding channel	对应信道(100)
a plurality of channels	多个通道(100)
a second counting channel	第二计数通道(100)
absolute grant channel	绝对许可信道(100)

access channel	1. 接入信道(78.16)
	2. 访问信道(17.84)
	3.存取信道(3.38)
	4.进入通道(0.56)
	5.出入通道(0.04)

Table2 Fuzzy Search Results (source: PatentLex)

2.7.2 Cytidine

The fuzzy search results of the multi-word term “cytidine” from two sources are provided for comparison below: PatentLex and HOWNET, a well-known Chinese language resource.

Cytidine
胞苷(91) 胞嘧啶核苷(16)
(source: HOWNET)

English Matched Term	Chinese Renditions (%)
cytidine	胞苷(100)

5-azacytidine	氮杂胞(61.79)
	5-氮杂胞(25.84)
	5-氮胞苷(12.35)

cytidine deaminase	胞苷脱氨酶(100)
cytidine monophosphate	胞苷一磷酸(100)
cytidine nucleotides	胞苷核苷酸(100)
cytidine triphosphate	胞苷三磷酸(100)
deoxy cytidine	脱氧胞苷(100)
deoxycytidine	脱氧胞苷(100)

Table 3 Fuzzy Search Results Compared (source: PatentLex)

In Table 4 below, authentic examples of usage from different domains of patents according to PCT classifications are provided from PatentLex, but not from HOWNET, because they are not available.

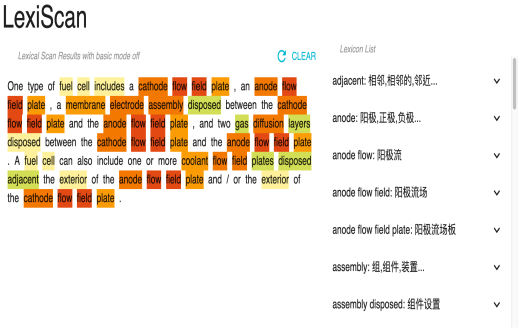
IPC	English	Chinese
C07	GDMEM contains DMEM (Gibco) and 4.5g/l glucose, 15 mg/1 phenol red, 1 mM sodium pyruvate, 1.75 g/1 sodium bicarbonate, 500 μM asparagine, 30 μM adenosine, 30 μM guanosine, 30 μM **cytidine**, 30 μM uridine, 10 μM thymidine, and non-essential amino acids (GIBCO).	GDMEM 培养基含 DMEM (Gibco)，4.5g/L 葡萄糖，15mg/L 酚红，1mM 丙酮酸钠，1.75g/L 碳酸氢钠，500 μm 天门冬酰胺，30 μm 腺苷，30 μm 鸟苷，30 μm **胞苷**，30 μm 尿苷，10 μ m 胸腺嘧啶核苷和非必需氨基酸 (GIBCO)。
A61	The cells were pre-incubated for 27.5 hours in **5-azacytidine** before addition of SAHA.	将细胞在 **5-氮杂胞苷**中预温育 27.5 小时，然后加入 SAHA 。
C07	The short plasma half-life is due to rapid inactivation of decitabine by deamination by liver **Cytidine deaminase**.	短的血浆半衰期是由于肝脏**胞苷脱氨酶**通过脱氨基作用对地西他滨的快速灭活而导致 。

Table 4 Some authentic example sentences for the alternate Chinese renditions (source: PatentLex)

It can be seen from the case of *cytidine* that there are considerable quantitative and qualitative differences between HOWNET and PatentLex. They show how the cultivation of a specialist corpus could expedite the user’s search and so enhance his productivity by optimizing his search.

2.8 Lexical Scan

Apart from being able to enjoy the useful feature of cross-lingual search engine which is helpful not just for translation, it is usual for a translator to face two challenges when tackling a new piece of text: (a) new technical terms not in his vocabulary, and (b) making a proper selection where there are multiple renditions. It would be very helpful if he or she is given some relative weighting (i.e., relative frequency of usage) of the alternate renditions, and if he or she could review actual examples from the technical texts where necessary. For this reason, a text scanning feature has been added by drawing on all bilingual multi-word expressions extracted in the previous effort. It includes (a) *renditions lookup*, (b) *distribution profiling*, (c) *authentic text lookup*, (d) *thesaurus navigation* in an integrated interface. An example of semantic net navigation of LexiScan involving (a) and (b) is given below:



LexiScan

Lexical Scan Results with basic mode off

One type of fuel cell includes a cathode flow field plate , an anode flow field plate , a membrane electrode assembly disposed between the cathode flow field plate and the anode flow field plate , and two gas diffusion layers disposed between the cathode flow field plate and the anode flow field plate . A fuel cell can also include one or more coolant flow field plates disposed adjacent the exterior of the anode flow field plate and / or the exterior of the cathode flow field plate .

Lexicon List

adjacent: 相邻,相邻的,邻近...

anode: 阳极,正极,负极...

anode flow: 阳极流

anode flow field: 阳极流场

anode flow field plate: 阳极流场板

assembly: 组,组件,装置...

assembly disposed: 组件设置

The LexiScan produces lexicon list which includes a full range of related terms in the sidebar:

adjacent: 相邻, 相邻的, 邻近
anode: 阳极, 正极
anode flow: 阳极流
anode flow filed: 阳极流场
anode flow filed plate: 阳极流场板
assembly: 组, 组件, 装置
assembly disposed: 组件装置

Given a piece of technical text, LexiScan will highlight all recognized technical terms from the PatentLex database, together with their renditions on the sidebar as shown (*renditions lookup*). Since a single word may be part of multiple multi-word expressions, a darker shade indicates that the single word contributes to other multi-word expressions. In the above LexiScan example, if we click on the word "*flow*" in the phrase "*cathode flow field plate*" on the first line, we can view the following list of words containing the word "flow" in English but not necessarily with the same renditions in Chinese.

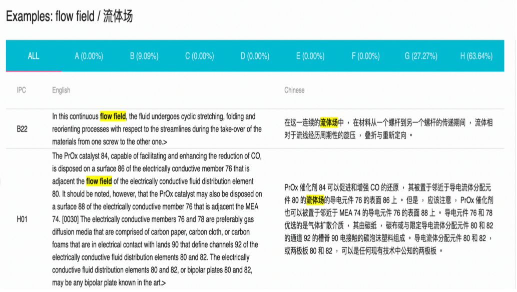
a. cathode flow: 阴极流
b. cathode flow field: 阴极流场
c. cathode flow field plate: 阴极流场板
d. flow field: 流场, 气流场, 流动区
e. flow field plate: 流场板

From the above, we can thus see a number of Chinese terms providing different renditions of "flow": namely "*cathode flow*", "*cathode flow field*", "*cathode flow field plate*", "*flow field*", and "*flow field plate*". They are all possible multi-word expressions covering "flow" from the lexicon. We could further expand some of the expressions on the list to examine the distribution of the possible renditions (*distribution profiling*).

Lexicon list (filtered) for flow field

a. 流场（94.94%）
b. 气流场（2.04%）
c. 流动区（1.69%）
d. 流动场（0.95%）
e. 流体场（0.35%）
f. 流动域（0.03%）

Furthermore, useful authentic bilingual example sentences drawn from patents may be obtained by clicking on "flow field" and "流体场" (*authentic text lookup*).



Examples: flow field / 流体场

IPC	English	Chinese
B22	In this continuous flow field, the fluid undergoes cyclic stretching, folding and reorienting processes with respect to the streamlines during the take-over of the materials from one screw to the other one.>	在这一连续的流体场中，在材料从一个螺杆到另一个螺杆的传递期间，流体相对于流线经历周期性的旋压，叠折与重新定向。
H01	The PrOx catalyst 84, capable of facilitating and enhancing the reduction of CO, is disposed on a surface 86 of the electrically conductive member 76 that is adjacent the flow field of the electrically conductive fluid distribution element 80. It should be noted, however, that the PrOx catalyst may also be disposed on a surface 88 of the electrically conductive member 76 that is adjacent the MEA 74. [0030] The electrically conductive members 76 and 78 are preferably gas diffusion media that are comprised of carbon paper, carbon cloth, or carbon foams that are in electrical contact with lands 90 that define channels 92 of the electrically conductive fluid distribution elements 80 and 82. The electrically conductive fluid distribution elements 80 and 82, or bipolar plates 80 and 82, may be any bipolar plate known in the art.>	PrOx 催化剂 84 可以促进和增强 CO 的还原，其被置于邻近于导电流体分配元件 80 的流体场的导电元件 76 的表面 86 上。但是，应该注意，PrOx 催化剂也可以被置于邻近于 MEA 74 的导电元件 76 的表面 88 上。导电元件 76 和 78 优选的是气体扩散介质，其由碳纸，碳布或与限定导电流体分配元件 80 和 82 的通道 92 的槽脊 90 电接触的碳泡沫塑料组成。导电流体分配元件 80 和 82，或两极板 80 和 82，可以是任何现有技术中公知的两极板。

As in cross-lingual search, the percentage distribution across patent domains can be seen from the headings. Bilingual examples belonging to a specific patent class can similarly be narrowed down by clicking on the heading. It is worth noting that the LexiScan feature has wider application in cross-lingual translation, especially for textual analytics.

These are recognized as useful provisions for professional translator as well as for students and teachers of translation.

2.9 Other Features Analysis

Additionally, there are wider applications, especially for cross-language patent search and analysis via the production of knowledge graphs which could help the user to navigate through relevant research networks (Tsou et al., 2019). Moreover, an intermediate stage during the 10 years long curation process has included the harvesting of a very sizable corpus of bilingually aligned sentence pairs useful for training and evaluating Chinese-English machine translation engines (Lu et al., 2011b; Goto et al., 2012, 2013; Tsou et al., 2017a, 2018).

2.10 Concluding Remarks

In the Age of Big Data, there is easy availability of data, but this ease in availability should not be mistaken for ease in securing quality. Considerable care and efforts must go into the cultivation and evaluation of the data for its full value to be realized. This paper has discussed how a database of 1 million entries of highly valued bilingual multi-word expressions in the technical fields has been profitably mined from a combined database involving 10 years of English and Chinese patents, comprising 5,353 million English words and 12,000 million Chinese characters. The processes involved and the utilization of the resultant database and platform, PATENTLEX, have been outlined. It is hoped that the case reported here has demonstrated the considerable value of well curated corpus not just for mining lexical gems, but for other data mining as well.

References

Gale, William A., and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In Proceedings of ACL. pp.79-85.

Dyer, Chris, Chahuneau, Victor and Noah A. Smith.: A simple, fast, and effective reparameterization of IBM Model 2. Proceedings of NAACL-HLT, pp. 644–648. (2013).

Adafre, Sisay Fissaha and Maarten de Rijke. 2006. Finding similar sentences across multiple languages in wikipedia. In Proceedings of EACL, pp. 62-69.

Brown, Peter F., Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. In Proceedings of ACL. pp.169-176.

Chen, Stanley F. 1993. Aligning sentences in bilingual corpora using lexical information. In Proceedings of ACL. pp. 9-16.

Goto, Isao, Bin Lu, Ka-Po Chow, Sumita Eiichiro, and Benjamin K. Tsou. (2012). "Overview of the Patent Translation Task at the NTCIR-9 Workshop". In Proceedings of the NTCIR-9 Workshop, pp. 559-578. Tokyo.

Goto, Isao, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou.: Overview of the patent machine translation task at the NTCIR-10 workshop. Proceedings of NTCIR-10 Workshop Meeting. (2013).

Och, Franz J., and Hermann Ney. 2004. The Alignment Template Approach to Machine Translation. Computational Linguistics, 30(4), 417-449.

Jiang Long, Shiquan Yang, Ming Zhou, Xiaohua Liu, and Qingsheng Zhu. 2009. Mining Bilingual Data from the Web with Adaptively Learnt Patterns. In Proceedings of ACL-IJCNLP. pp. 870-878.

Tian Liang, Fai Wong, and Sam Chao.: Phrase Oriented Word Alignment Method. In Wang, Hai Feng (Ed.), Proceedings of the 7th China Workshop on Machine Translation pp. 237‑250. Xiamen, China (2011).

Tian Liang, Derek F. Wong, Lidia S. Chao, and Francisco Oliveira.: A Relationship: Word Alignment, Phrase Table, and Translation Quality. The Scientific World Journal1–13 (2014).

Bin Lu, Benjamin K. Tsou, Jingbo Zhu, Tao Jiang, and Olivia Y. Kwong. (2009). The Construction of an English-Chinese Patent Parallel Corpus. MT Summit XII 3rd Workshop on Patent Translation.

Bin Lu, Benjamin K. Tsou, Tao Jiang, Oi Yee Kwong and Jingbo Zhu. 2010a. Mining Large-scale Parallel Corpora from Multilingual Patents: An English-Chinese example and its application to SMT. In Proceedings of the 1st CIPS-SIGHAN Joint Conference on Chinese Language Processing (CLP-2010). Beijing, China.

Bin Lu, Ka-po Chow and Benjamin K. Tsou. 2011a. The Cultivation of a Trilingual Chinese-English-Japanese Parallel Corpus from Comparable Patents. In *Proceedings of Machine Translation Summit XIII (MT Summit-XIII)*. Xiamen.

Bin Lu, Benjamin K. Tsou, Tao Jiang, Jingbo Zhu, and Kwong, Olivia. 2011b. "Mining parallel knowledge from comparable patents". In Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances. pp. 247-271. IGI Global.

Bin Lu, Ka-po Chow and Benjamin K. Tsou (2015). "Comparable Multilingual Patents as Large-scale Parallel Corpora". In: Serge Sharoff, Reinhard Rapp, Pierre Zweigenbaum, Pascale Fung (Eds.), Building and Using Comparable Corpora. Springer-Verlag, pp 167-187.

Bin Lu, Benjamin K. Tsou and Ka-po Chow. (2016). Cultivating Large-scale Parallel Corpora from Comparable Patents: From Bilingual to Trilingual, and Beyond. In Tsou, Benjamin, and Kwong, Olivia., (eds.), Linguistic Corpus and Corpus Linguistics in the Chinese Context (Journal of Chinese Linguistics Monograph Series No.25). Hong Kong: The Chinese University Press, pp. 447-471.

Xiao-yi Ma.: Champollion: A Robust Parallel Text Sentence Aligner. In Proceedings of the 5th

International Conference on Language Resources and Evaluation (LREC). Genova, Italy (2006).

Utiyama, Masao, and Isahara Hitoshi. 2007. A Japanese-English patent parallel corpus. In Proceeding of MT Summit XI. pp. 475–482.

Utiyama, Masao and Isahara Hitoshi.: Reliable measures for aligning Japanese-English news articles and sentences. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pp. 72– 79. Sapporo, Japan (2003).

Simard, Michel, and Plamondon Pierre. 1998. Bilingual Sentence Alignment: Balancing Robustness and Accuracy. Machine Translation, 13(1), 59-80.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In Proceedings of MT Summit X.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, et al. 2007. Moses: Open source toolkit for statistical machine translation. In Proceedings of ACL Demo Session. pp. 177-180.

Koehn, Philipp.: Statistical machine translation. Cambridge University Press, the United Kingdom (2010).

Resnik, Philip and Smith Noah A. (2003) The Web as a Parallel Corpus, Computational Linguistics 2003 29:3, pp. 349-380

Haldar, Rishin and Mukhopadhyay Debajyoti. "Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach." ArXiv abs/1101.1232 (2011).

Jason R. Smith & Quirk, Chris & Toutanova, Kristina. (2010). Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. 403-411.

Sharoff S., Rapp R., Zweigenbaum P. (2013) Overviewing Important Aspects of the Last Twenty Years of Research in Comparable Corpora. In: Sharoff S., Rapp R., Zweigenbaum P., Fung P. (eds) Building and Using Comparable Corpora. Springer, Berlin, Heidelberg.

Smith, Jason R., Chris Quirk and Kristina Toutanova. 2010. Extracting Parallel Sentences from Comparable Corpora using Document Level Alignment. In Proceedings of NAACL-HLT. pp. 403-411.

Benjamin K. Tsou, Bi-wei Pan, and Ka-po Chow. Some Challenges and Advances in Natural Language Processing of Chinese Patents ─ From Machine Translation to Cognitive Filtering. [Keynote paper], WIPO East Meets West Seminar. Vienna (2017a).

Benjamin K. Tsou, Derek Wong, and Ka-po Chow. Successful Generation of Bilingual Chinese-English Multi-word Expressions from Large Scale Parallel Corpora: An Experimental Approach, paper presented at EUROPHRAS. London (November 2017b).

Benjamin K. Tsou, Min-yu Zhao, Bi-wei Pan, and Ka-po Chow.: The Age of Big Data and AI: Challenges and Opportunities for Technical Translation 4.0 and Relevant Training. Translators Association of China (TAC) Conference. Beijing (2018).

Benjamin K. Tsou, Ka-po Chow, Jun-ru Nie and Yuan Yuan: Toward a Proactive MWE Terminological Platform for Cross-Lingual Mediators in the Age of Big Data. Second Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT) (2019).

Germann, U. Aligned Hansards of the 36th Parliament of Canada (2001), http://www.isi.edu/natural-language/download/hansard/

Dekai Wu, and Pascale Fung 2005. Inversion transduction grammar constraints for mining parallel sentences from quasi-comparable corpora. In Proceedings of IJCNLP2005.

Xiaodong Zeng, Lidia S. Chao, Derek F. Wong, Isabel Trancoso, and Liang Tian.: To- ward Better Chinese Word Segmentation for SMT via Bilingual Constraints. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Vol. 1), pp. 1360–1369. Baltimore, Maryland (2014).

Bing Zhao, and Stephen Vogel. 2002. Adaptive Parallel Sentences Mining from Web Bilingual News Collection. In Proceedings of Second IEEE International Conference on Data Mining (ICDM-02).