Towards a Proactive MWE Terminological Platform for Cross-Lingual Mediation in the Age of Big Data

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

Junru Nie
Hong Kong University of Science and Technology
Chilin (HK) Ltd
ulricanie@gmail.com

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

Yuan Yuan
Hong Kong University of Science and Technology
Chilin (HK) Ltd
belleyuan26@gmail.com

Abstract

The emergence of China as a global economic power in the 21st Century has brought about surging needs for cross-lingual and cross-cultural mediation, typically performed by translators. Advances in Artificial Intelligence and Language Engineering have been bolstered by Machine learning and suitable Big Data cultivation. They have helped to meet some of the translator's needs, though the technical specialists have not kept pace with the practical and expanding requirements in language mediation. One major technical and linguistic hurdle involves words outside the vocabulary of the translator or the lexical database he/she consults, especially Multi-Word Expressions (Compound Words) in technical subjects. A further problem lies in the multiplicity of renditions of a term in the target language.

This paper discusses a proactive approach following the successful extraction and application of sizable bilingual Multi-Word Expressions (Compound Words) for language mediation in technical subjects, which do not fall within the expertise of typical translators, who have inadequate appreciation of the range of new technical tools available to help him/her. Our approach draws on the personal reflections of translators and teachers of translation and is based on the prior R&D efforts relating to 300,000 comparable Chinese-English patents. The subsequent protocol we have developed aims to be proactive in meeting four identified practical challenges in technical translation (e.g. patents). It has broader economic implication in the Age of Big Data (Tsou et al, 2015) and Trade War, as the workload, if not, the challenges, increasingly cannot be met by currently available front-line translators. We shall demonstrate how new tools can be harnessed to spearhead the application of language technology not only in language mediation but also in the "teaching" and "learning" of translation. It shows how a better appreciation of their needs may enhance the contributions of the technical specialists, and thus enhance the resultant synergetic benefits.

1 Two Converging Paths in Cross-Language Mediation

Translation and cross-lingual mediation are no longer exclusively human efforts but draw on many indispensable tools and resources which have resulted from successful and fruitful research and development efforts in natural language processing (Bowker and Pastor, 2015). We highlight four major stages in the translator's workflow, in which distinct technical efforts could enhance productivity (Zaretskaya et al., 2015).

1.1 From the Perspective of Translators

The translator's workflow consists of four major stages. When working with a technical document, even if he/she has excellent command of the languages concerned, it is inevitable that there will be unfamiliar terms outside his/her active vocabulary.

Temnikova, I, Orăsan, C., Corpas Pastor, G., and Mitkov, R. (eds.) (2019) *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, Varna, Bulgaria, September 5 - 6, pages 116–121.
https://doi.org/10.26615/issn.2683-0078.2019_014

A. To cope with these challenges, appropriate lexical resources and other reference materials have to be consulted. Therefore, he/she needs to have convenient access to useful and easily manageable databases. The major challenge is the *Accessibility* of suitable reference materials.

B. Quite often dictionaries provide multiple renditions of given terms appropriate to only some appropriate domains. He/she has to adjust his/her selection for the translation task at hand. The major hurdle at this stage is *Adjustability* in selecting the suitable subset of renditions within the right domain.

C. Having access to the multiple renditions is not sufficient, and access to authentic examples on the use of the alternate renditions would be helpful for making his/her selection. The issue of *Accountability* of the lexical variations is a major requirement at this stage.

D. For self-improvement, the conscientious translator or the student of translation would find it useful to be able to browse through a new relevant lexical database in serendipity search so as to uncover related and associated terms and renditions. This may be seen as a desirable feature of *Adaptability* of the database whereby the user may advance his/her lexical knowledge.

1.2 From the Perspective of the Computer Scientist

To help to cope with the four A issues: ***Accessibly, Adjustability, Accountability*** and ***Adaptability*** concerning the lexical hurdles of the translator, the computer scientist's concern would be to provide a suitable database which would contain the relevant terms and translation tools for the translator. He/she would need to focus on several distinct tasks (Sections 1.1 and 1.2 are cross-referenced):

A. To **secure** the best database in order to produce the best lexical resources for the translator. He/she would be concerned with the identification and access a suitable textual corpus and the use of the best algorithms to accomplish the matching of bilingual terms. Objective indices such as Precision and Recall, F measures which are purely statistically based, would be upmost on his/her mind (Mitkov, 2016, 2017). As he/she is in most cases unlikely to be knowledgeable with wide-ranging linguistic issues in both languages, he/she would be using the *"Happy Majority Approach"* whereby meeting the statistically significant requirements of the majority would be happily acceptable under normal circumstances. The professional translator demands much more just as his/her demands are incrementally met.

B. The ideal one-to-one matching of the terms and their meanings fall by the wayside very readily and the computer scientist has to deal with the "one-to-how many" problems. It is a major challenge to determine the full range of alternate target renditions and to uncover and select the subset of the results to suit the needs of the users. For example, a common term "multiplication" in arithmetic refers to specifically the number of times an item or a sum is replicated (乘法). However, in biological sciences, it refers to reproductive generation (繁殖，衍生) without the precision required in arithmetic, and must be translated accordingly. The average individual would have the arithmetic sense foregrounded in his/her mind, and only when bilingual texts in English and Chinese are contrasted would the additional sense of reproduction be likely brought to mind. This provision of the multiple alternate renditions is very much appreciated by the translators.

C. In the longer term a necessary feature would be an updated database of terms with representative authentic examples from authoritative technical document (Lu et al., 2011). Such a database would be welcome by the translators as a dynamically maintained thesaurus.

D. The provision of knowledge graph and semantic network on the basis of large textual databases has made considerable advances.

E. It is especially useful for the translator and language mediator who, for self-improvement, is keen to search beyond a single target word to explore related and associated words.

2 Pairing Cross-Lingual Terms

Based on the bilingual MWE database, we have constructed a cross-lingual search MWE platform – PatentLex (Tsou et al., 2017). 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 searchquery function, as can be seen in Table1.

Matched Term (English)	Renditions (Chinese)
heat pump	1. 热泵(98.97%) heat-pump 2. 加热泵(0.67%) add-heat-pump 3. 供热泵(0.28%) supply-heat-pump 4. 受热泵(0.07%) receive-heat-pump
absorption heat pump	吸收式热泵(100%)
air conditioners and heat pumps	1. 空调和热泵(66.66%) 2. 空气调节器和热泵(33.33%)
bernoulli heat pump	1. 柏努利热泵(59.25%) 2. 伯努利热泵(40.74%)
bernoulli heat pumps	伯努利热泵(100%)
chemical heat pump	化学热泵(100%)
chemical heat pumps	化学热泵(100%)
conventional heat pumps	常规热泵(100%)

Table 1: Multiple Chinese renditions of *Heat Pump*.

2.1 "Heat Pump"

Of the four possible renditions: "热泵" (heat-pump), "加热泵" (add-heat-pump), "供热泵" (supply-heat-pump) and "受热泵" (receive heat pump), it is noteworthy that some of these Chinese renditions are more informative than the English term. For example, heat pump in English has been rendered as "加热泵" (add-heat-pump) which is a better rendition as it indicates one function of the heat pump in Table 2 below.

No.	IPC[1]	English	Chinese
1	C09	While the primary purpose of refrigeration is to remove energy at low temperature, the primary purpose of a **heat pump** is to add energy at higher temperature.	致冷的首要目的是在低温时除去能量，而**热泵**的首要目的是在高温时增加能量。
2	H02	The potential benefits include one or more of reduced air noise, better dehumidification, warmer air in **heat pump** mode, or the like.	其潜在益处包括下列的一种或几种，即减小的噪音、更好的除湿、**加热泵**模式中温热的空气或类似情况。

Table 2: Authentic examples.

The advantages of these optional details are two-fold: they provide a rudimentary semantic network of associated concepts of the original target terms, and they also alert the translators that the search term may have other possible renditions when considered in a larger context.

2.2 "Wafer"

In the Table 3 below, a comparison is made between the provisions made by a well-known Chinese language resource: HOWNET (http://dict.cnki.net/dict_result.aspx), and by PatentLex. HOWNET's source data is not limited to technical documents, and their bilingual search engine also provides different renditions with information on relative frequencies, though not according to domains.

[1] IPC: International Patent Classification.

PatentLex	HOWNET
1. 晶片(95.29%) 2. 硅片(2.9%) 3. 圆片(1.53%) 4. 晶圆(0.13%) 5. 糯米纸(0.07%) 6. 薄脆饼(0.06%)	1. 晶片（32.65%） 2. 硅片（58.73%） 3. 干胶片（0%） 4. 圆片（8.63%）

Table 3: Alternate renditions of *Wafer* in Chinese and English.

It can be seen that both HOWNET and PatentLex offer alternate renditions of this technical term. However, PatentLex offers 2 more renditions than HOWNET. Furthermore, HOWNET's third rendition shows "干胶片" with 0% of usage, whereas it is not found in PatentLex's technical literature. PatentLex's "晶片"(95.29%) is the top choice in Patentlex whereas the top choice item from HOWNET "硅片"(58.73%) has only 2.9% usage in the technical texts represented by PatentLex. The broader search results of the term *Wafer* are as follows.

Matched Term（English）	Renditions (Chinese)
1.wafer	1. 晶片(95.29%) 2. 硅片(2.9%) 3. 圆片(1.53%) 4. 晶圆(0.13%) 5. 糯米纸(0.07%) 6. 薄脆饼(0.06%)
2.adjacent wafers	1. 相邻晶片(72.97%) 2. 相邻板片(27.02%)
3.bare silicon wafer	裸硅晶片(100%)
4.bonded wafers	键合的晶片(100%)
5.bottom side of the wafer	晶片底面(100%)
6.applied to the wafer	1. 施加到晶片(87.17%) 2. 应用到晶片(12.82%)
7.attached to the wafer	附着到晶片(100%)
8.backside wafer pressure	背面的晶片压力(100%)

Table 4：Fuzzy search of *Wafer.*

Some authentic examples from a wide range of alternative renditions are given in Table5.

No.	IPC[2]	English	Chinese
1	H01	Therefore, a center of rotation of the semiconductor **wafer** W can be kept in a constant position.	因此，半导体**晶片** W 的旋转中心可以被保持在恒定位置。
2	C08	The water droplet contact angle was measured within 2 or 3 seconds of placing the droplet 64 on the coated **wafer** surface.	在 2 或 3 秒内测量放置于涂布**硅片**表面上的水滴接触角。
3	A61	The implant 102 is preferably formed of relatively thin **wafer** of biologically compatible material.	植入物 102 较佳地由生物学上相容的材料做成的相当薄的**圆片**形成。
4	C07	The compound of Formula (I) can also be incorporated into a candy, a **wafer**, and/or tongue tape formulation for administration as a "quick-dissolve" medicament.	还可以将式 (I) 的化合物掺入到糖果、**糯米纸**和 / 或舌粘带制剂以 " 速溶 " 药物的形式给药。
5	A21	The present invention therefore addresses the problem of how to provide an approximately circular **wafer** which also has the desired crispness.	因此，本发明致力于如何提供一种大体上为圆形的、同时又具备理想的松脆度的**薄脆饼**制品的问题。

Table 5: Authentic examples of alternate renditions.

It may be noted that 糯米纸 "glutinous rice paper" (No.5 Table4) and 薄脆饼 "thin crisp cake"(No.6 Table4) are generally not technical but culinary terms. Nonetheless they can be found in the

[2] IPC: International Patent Classification.

technical database of PatentLex under medical sciences (C07, Table5) and food industry (A21, Table5) respectively rather than just in a general language resource database.

3 Lexi Scanning

Prior to being able to access alternate renditions of a given technical term, the translator is confronted by the related and practical problem of encountering words which are altogether out of his/her vocabulary. Thus, a platform through which a translator may submit a text he/she has to work on and which could provide indications of all the embedded terms in the database through highlighting would be very much welcome. Such a provision is made by PatentLex with 1 million entries of preloaded bilingual MWE's (Tian et al., 2014; Tsou et al., 2018, 2019). The process of derivation also produced parallel sentences useful for MT research and MT evaluation (Goto et al., 2012, 2013).

4. Mining Knowledge Graph

We can construct a knowledge graph based on the bilingual term database, together with the details of the distribution of the alternate renditions. This makes use of dynamic information drawn from authentic patent documents and compiled statistics, rather than static information as found in ordinary dictionaries or handcrafted web of semantic terms. This reflects real world usage and also enables knowledge map navigation through the links between different terms and concepts.

For example, from "channel", we can obtain a list of possible related renditions in both languages with their relative frequencies, as illustrated in the chart below. If we click on a target English word "channel"(1), we will be led to 6 Chinese renditions (a) 通道(10.92%); (b) 途徑(0.02%); (c) 頻道 (3.06%); (d) 路線(0.01%); (e) 槽(2.55%); (f) 信道 (30.89%), each with its frequency of usage indicated. If we then choose one of the Chinese nodes, for example, (a)通道(10.92%), we are led to 5 other English terms besides the original relevant "channel" such as (2) aisle (0.04%); (3)passageway (0.49%); (4) access tunnel (0.4%); (5) conduit (6.86%); (6) passage (17.22%). We could proceed further by clicking on one of term such as (5) conduit(6.86%), and three Chinese actual renditions will be indicated: (g)導 管(5.88%); (h)管 線 (2.72%); (i)管道(41.36%). This dynamic thesaurus would facilitate the work of the protocol user. If we choose to search more deeply by following one of the renditions, such as "通道", we will obtain another set of renditions and percentages. Likewise, we can drill deeper and navigate along the rendition "conduit" and uncover another set of 3 renditions "導管", "管線" and "管道". The flowchart below illustrates the paths of navigation.

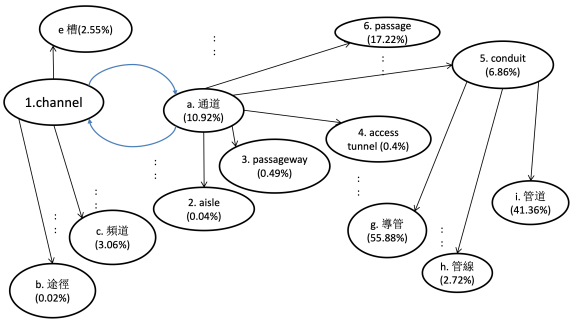

Figure 1: Flow chart: "Channel" vs "通道" bilingual knowledge graph navigation.

The provision of bilingual-knowledge graphs such as represented by the above flow chart would be useful for multilingual as well as monolingual searches.

5. Concluding Remarks

In the Age of Big Data, there is easy availability of data for developing resources and tools for translators and cross-language mediation (Tsou, 2018, 2019; Tsou et al., 2018). Four key stages in the workflow of translators have been identified with four overlapping areas in mature and developing languages technology. On the basis of an expanding database of more than one million entries of highly valued bilingual multi-word expressions in the technical fields we have developed a bilingual MWE platform, which shows how an articulated protocol could be organized proactively for translators with purposeful utilization of NLP results and tools. (Tsou et al, 2019) While some of the features are found in existing tools such as Trados (https://www.sdl.com/software-and-services/translation-software/terminology-management/sdl-multiterm/) and HOWNET, for example, Patentlex has attempted to incorporate all of them into a single platform. It is hoped that the welcomed coordinated approach underlying the PatentLex platform will allow similar efforts to be attempted for other language pairs.

References

Guihong Cao, Jianfeng Gao and Jianyun Nie. 2007. A System to Mine Large-scale Bilingual Dictionaries from Monolingual Web Pages. In Proceedings of MT Summit, pages 57-64.

Chiang, David. 2007. *Hierarchical phrase-based translation. Computational Linguistics*, 33(2), pages 201–228.

Fujii, Atsushi, Masao Utiyama, Mikio Yamamoto, and Takehito Utsuro. 2008. Overview of the patent translation task at the NTCIR-7 workshop. In Proceedings of the NTCIR-7 Workshop, pages 389-400. Tokyo, Japan.

Fujii, Atsushi, Masao Utiyama, Mikio Yamamoto, Takehito Utsuro, Terumasa Ehara, Hiroshi Echizen-ya and Sayori Shimohata. 2010. Overview of the patent translation task at the NTCIR-8 workshop. In Proceedings of the NTCIR-8 Workshop. Tokyo, Japan.

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

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

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

Ruslan, Mitkov. 2016. "Computational Phraseology light: automatic translation of multiword expressions without translation resources." Yearbook of Phraseology 7.1 pages 149-166.

Ruslan, Mitkov. 2017. "Computational and Corpus-Based Phraseology: Second International Conference", Europhras 2017, London, UK, November 13-14, Proceedings. Vol. 10596. Springer.

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

Liang Tian, Derek F. Wong, Lidia S. Chao, and Francisco Oliveira. 2014. *A Relationship: Word Alignment, Phrase Table, and Translation Quality. The Scientific World Journal,* pages 1–13.

Benjamin K. Tsou. 2018. Patent Translation and Text Analysis: Opportunities and Challenges in the Age of Big Data. 9TH CHINA PATENT ANNUAL CONFERENCE. Beijing.

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

Benjamin K. Tsou, Min-yu Zhao, Bi-wei Pan, and Ka-po Chow. 2018. 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.

Benjamin K. Tsou and Kapo Chow. 2019. From the cultivation of comparable corpora to harvesting from them: A quantitative and qualitative exploration. Proceedings of the 12th Workshop on Building and Using Comparable Corpora. Varna, Bulgaria.

Benjamin K. Tsou and Olivia Kwong. 2015. LIVAC as a Monitoring Corpus for Tracking Trends beyond Linguistics. 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, pages 447-471.

Dekai Wu, and Xuanyin Xia. 1994. Learning an English-Chinese lexicon from a parallel corpus, In Proceedings of the First Conference of the Association for Machine Translation in the Americas.

Anna Zaretskaya, Gloria Corpas Pastor and Miriam Seghiri. 2015. "Translators' requirements for translation technologies: A user survey." New Horizons in Translation and Interpreting Studies, pages 133-134.

Lynne Bowker and Gloria Corpas Pastor. 2015. "Translation technology." *The Oxford Handbook of Computational Linguistics 2nd edition*.