From Fidelity to Fluency: Natural Language Processing for Translator Training

Oi Yee Kwong
Department of Translation
The Chinese University of Hong Kong
oykwong@arts.cuhk.edu.hk

Abstract

This study explores the use of natural language processing techniques to enhance bilingual lexical access beyond simple equivalents, to enable translators to navigate along a wider cross-lingual lexical space and more examples showing different translation strategies, which is essential for them to learn to produce not only faithful but also fluent translations.

1 Introduction

Online dictionaries are important computer-aided tools for translators today (Bowker, 2015), while parallel corpora, despite their relative scarcity, have become useful resources for translation teaching (Olohan, 2004). The two kinds of reference provide what lexicographers like Atkins and Rundell (2008) would distinguish as context-free and context-sensitive translations respectively. The current work, as a prelude to a larger project, discusses the limitations of existing bilingual lexical resources and proposes natural language processing approaches for enhancing their navigational means for better usability in translator training and computer-aided translation.

Consider the translation of the English sentence "I still have vivid memories of that evening" into Chinese. The Online Cambridge English-Chinese Dictionary[1] shows two senses of "vivid", and quite straightforwardly the word can be disambiguated between the first sense (Vivid descriptions, memories, etc. produce very clear, powerful, and detailed images in the mind) and the second sense (very brightly coloured). Hence, notwithstanding the normal associative strengths between words, when "vivid" has been properly disambiguated, its associations with "colour", "bright", etc. are down-weighted compared with its associations with "recollection", "memory", "clear", etc.

Once the decoding purpose is fulfilled, with the appropriate senses identified ("vivid" as above and "memory" as "something that you remember from the past"), one can then refer to the Chinese "equivalents" provided by the dictionary: 栩栩如生的, 鮮活的, and 生動的 for "vivid", and 記憶 and 回憶 for "memory". But the encoding purpose is not achieved yet, because none of the combinations between these lexical items could be considered satisfactory. They are only conceptually close to what we need, but not exactly appropriate for the context. It will only be helpful if we can depart from them and navigate further along their associations. The ability to do so is essentially what translator training would need to foreground, especially for novice translators to produce not only faithful but also fluent translations.

In the rest of this paper, we will first illustrate, in Section 2, the limitations of existing bilingual resources from the cognitive perspective, especially with reference to word associations. We will discuss in Section 3 the implications on the need for enhanced access of those resources to facilitate translator training. Section 4 outlines the natural language processing techniques employed in our ongoing work in this regard.

2 Word Association for Lexical Access

Word association has been deemed an important element in the mental lexicon (e.g. Collins and Loftus, 1975; Aitchison, 2003) as well as many lexical resources employed in a variety of natural language processing tasks (e.g. Fellbaum, 1998; Navigli and Ponzetto, 2012), and is believed to be able to provide useful navigational means to address the search problem in lexical access in dictionaries (Zock et al., 2010).

[1] http://dictionary.cambridge.org

Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications, pages 130–134
Melbourne, Australia, July 19, 2018.

While there are various ways to model different associative relations from large corpora (e.g. Church and Hanks, 1990; Wettler and Rapp, 1993; Biemann et al., 2004; Kilgarriff et al., 2004; Hill et al., 2015), certain knots remain to be untied for them to be better utilised in language applications. First, corpus-based modelling of associations often focuses on specific relations (e.g. similarity, hierarchical relations, collocations, etc.), but in real-life lexical access, a combination of relations is often retrieved, as shown in human word association norms (e.g. Moss and Older, 1996). Moreover, some associations are bound to be more relevant than others in a given context, and they are readily activated regardless of their normal associative strengths. Second, for tasks requiring bilingual lexical access, care must be taken especially when onsidering the non-identical conceptual and linguistic structures across languages. Given the scarcity of complete equivalence and different linguistic properties, bilingual (or multilingual) word associations based entirely on bi- or multi-lingual concept lexicalisations (equivalents) may not be adequate for representing the cross-lingual word association patterns.

Existing bilingual dictionaries nevertheless generally presume the existence of lexical translation equivalents. Analysis of human association responses, as in Kwong (2013; 2016), suggests an alternative view. On the one hand, very different association types are found for different word classes (e.g. more taxonomic associations for nouns and more collocational associations for verbs), and across English and Chinese (e.g. more paradigmatic responses for English but clear preference for syntagmatic associations for Chinese). On the other hand, free associations may be modelled from large corpora, but the results vary considerably for individual words, some even counter-intuitive. Less frequent associations are normally disadvantaged, but humans readily retrieve them when prompted by a certain context. Hence, modelling of associations should be task-driven.

In addition, the equivalents given in bilingual lexicons are basically decontextualized, and they often do not appear in the example bilingual sentences in the dictionaries. Thus, an association found in the source language may not hold for the equivalents found in a target language. When using word associations in a bilingual context, other than associative strengths, cross-lingual correspondence of the associations is also worth investigation.

One conventional issue in psycholinguistics regarding models of bilingual lexicon is whether the conceptual stores for two languages are shared or separated (Keatley, 1992), and many studies suggest that the store is mostly shared (e.g. Kroll and Sunderman, 2003). Another issue is what is shared and what is separated in particular lexical concepts (Jarvis and Pavlenko, 2008). Pavlenko (2009) suggested, in contrast to the conclusions by many, that weaker connections failing to show a semantic priming effect may not necessarily indicate the lack of shared meaning, as conceptual equivalence can range from complete equivalence to partial and even non-equivalence, and the bilingual mental lexicon undergoes conceptual restructuring during language learning when cross-linguistic differences are encountered. Such cognitive aspects may not have been sufficiently modelled in static bilingual linguistic lexicons, especially between two very different languages like English and Chinese.

In the following we will compare the word associations obtained from various resources, and evaluate them against the information need in our earlier example situated in the translation context.

2.1 Word Association Norms

Table 1 shows the non-single responses in descending order of frequency in the University of South Florida (USF) Association Norms (Nelson et al., 1998), for the stimuli "vivid" and "memories". Apparently, should "vivid" and "memories" be associated, they are linked by "dream". In fact, "memory" was among the 33 single responses for "vivid", while "vivid" was not among any of the responses for "memories" or "memory".

vivid	**memories**	
clear	past	album
color	thoughts	cats
bright	happy	**good**
imagination	pictures	love
real	**dreams**	photos
alive	mind	tears
dream	**bad**	boyfriends
read	childhood	fond
unclear	friends	high school
natural	remember	recollections
strong	songs	sad

Table 1: Responses from USF Association Norms

The equivalents in the Online Cambridge Dictionary for "vivid" (栩栩如生(的), 鮮活(的), and

清晰 (clear)	**印象 (impression)**
可見 (visible)	深刻 (deep)
目標 (objective)	印象派 (impressionism)
指引 (guideline)	**良好 (good)**
模糊 (unclear)	**差 (bad)**
清楚 (clear)	人 (person)
影像 (image)	第一印象 (first impression)
明白 (understand)	派 (-ism)

Table 2: Responses from HKC Association Norms

生動(的)) and for "memory" (記憶 and 回憶) are not found in the Hong Kong Chinese (HKC) association norms (Kwong, 2013), so instead we look at the responses for two similar items, 清晰 (clear/vivid) and 印象 (impression/memory), respectively[2]. The non-single responses for these stimuli are shown in Table 2. For 清晰, the responses 清楚 (clear) and 模糊 (unclear) can be said to match the English responses for "vivid", but other than that the response patterns differ considerably across languages. The only response related to "memory" is 印象 which appeared only once. Similarly, the stimulus 印象 has its own cluster of associations and the most typical adjective associated with it (深刻) is not one expected in English for "memories", although more general ones like "good" and "bad" are found in common.

2.2 Dictionary Text

Based on the content words gathered from the definitions in the Online Cambridge English-Chinese Dictionary (Table 3), it seems that "vivid" and "memories" are closely associated, with the latter appearing in the definition of the former. But as mentioned above, one cannot really take the given Chinese equivalents and combine them for the translation. None of the combinations would sound idiomatic to a native Chinese speaker.

vivid	**memory**
descriptions	something
memories	remember
produce	past
clear	
powerful	
detailed	
images	
mind	

Table 3: Associations from Dictionary Definitions

[2]The former is among the equivalents for "vivid" in iCIBA (http://www.iciba.com/) and the latter is a near-synonym for 記憶 in a Chinese dictionary (http://dict.revised.moe.edu.tw).

2.3 Large Corpora

Making use of the Word Sketch function for selected *gramrel* collocations and the Thesaurus function in the Sketch Engine (Kilgarriff et al., 2004; Rychlý and Kilgarriff, 2007) on the ukWaC corpus and twWaC corpus, Tables 4 and 5 show the top 10 results for our target words.

vivid		**memory**	
modifies	**thesaurus**	**modifier**	**thesaurus**
recollection	compelling	fond	image
imagination	vibrant	loving	thought
evocation	evocative	childhood	knowledge
imagery	poignant	short-term	picture
depiction	colourful	distant	feeling
memory	imaginative	**vivid**	sense
portrayal	striking	collective	vision
dream	fascinating	episodic	experience
color	dramatic	flash	character
portrait	memorable	happy	idea

Table 4: Associations from ukWaC

清晰		**回憶**	
noun_right	**thesaurus**	**adj_left**	**thesaurus**
影像	清楚	美好	美好
照片	模糊	共同	記憶
概念	完整	老	童年
畫面	生動	許多	回想
聲音	深刻	深刻	時光
條理	流暢	難忘	快樂
輪廓	鮮明	浪漫	故事
認識	明確	永久	感動
文字	簡單	不愉快	往事
方向	呈現	永生	難忘

Table 5: Associations from twWaC

The following are noted from the results. First, in English, "vivid" and "memory" are strongly collocated, as the same collocation pops up from both directions (what does "vivid" modify / what modifies "memory"). But to a certain extent, whether an expected association can be extracted depends on individual corpora. For instance, with thesaurus function on ukWaC, "recollection" (synonym of "memory") is not even found, and the near-synonym "impression" ranked after the 450th place. Second, very little overlap is found between the English and Chinese associations extracted (even if based on partial equivalents). Arguably we started with partial equivalents anyway (but that is inevitable), and it shows that the word association patterns may not be the same across translation equivalents.

3 Implications

Realising that Adj-N constructions in English are not necessarily rendered as Adj-(的de)-N in Chinese, one must go beyond the context-free equivalents given in bilingual dictionaries to look for potential target expressions which may sometimes be found from the context-sensitive translations shown in the example sentences. While one might faithfully combine the bilingual lexicalisations of "vivid" and "memory" to give 生動/鮮活/逼真/清晰的記憶, other more idiomatic and fluent ways of expressing the same meaning in Chinese should be accessible for reference, including word-class shifts like 清楚記得/記得清清楚楚 (remember vividly), use of four-character expressions like 記憶猶新, as well as other appropriate expressions depending on context, such as 印象深刻 and 歷歷在目, to name a few examples.

The process of determining the appropriate target expression from the partial equivalents can sometimes be tricky especially considering the word formation, polysemy, and collocation patterns across the two languages (e.g. even for the same sense, "clear" appropriately corresponds to 清晰/清楚 when collocated with image/explanation respectively, and 清澈/透明 with river/glass respectively). The challenge is even more pronounced when no correspondence can be spotted from the examples, or for generally weakly associated words (e.g. strong-endorsement). Thus, natural language processing techniques are adopted to enhance bilingual access beyond lexical equivalents for translators.

4 Work in Progress

It is not simply lexical transfer but a transfer of the whole relevant semantic space that is needed in translation. With this in mind, we are pursuing two routes using natural language processing approaches to enhance bilingual lexical access beyond simple translation equivalents, for reference in the translation process.

The first involves chaining up collocation information in a cross-lingual manner. Many have realised that there are often conceptual gaps across languages, but in addition to the bilingual correspondences of individual lexicalised concepts, it is necessary to consider the cross-lingual difference in terms of not only conceptual structure but also collocation patterns. As McKeown and Radev (2000) pointed out, a concept expressed by way of a collocation in one language may not have a corresponding collocation in another language.

Hence, ideally one should be able to start from a certain collocation or cluster of collocation in one language (e.g. vivid-memory) and, through some translation equivalents as seed words (e.g. memory-回憶), extend into the relevant semantic space in the other language (e.g. 往事/歷歷/印象/深刻) which is otherwise unretrievable from bilingual lexicons alone, as Figure 1 shows. For experiments, the Bilingual Word Sketch function in the Sketch Engine (Baisa et al., 2014) is taken as a starting point, upon which strategic application of word sense disambiguation, clustering, and word embedding techniques is tested for their effects on re-prioritising word associations with respect to specific collocations for a given context.

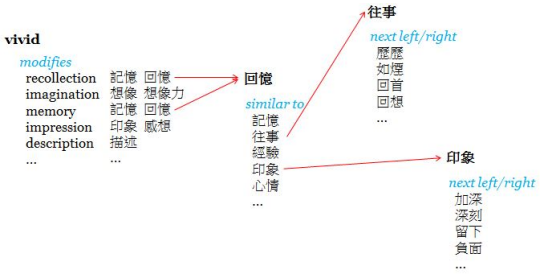

Figure 1: A Glimpse of a Cross-lingual Collocation Chain

The second makes use of neural machine translation (NMT) to obtain paraphrase sentence pairs. While most machine translation research focuses primarily on the fidelity of the target text, other possible and perhaps more fluent renditions are either ranked very low or completely ignored. They may exist in parallel corpora but with so low a frequency that often leaves NMT models to consider them noise. Thus we propose to identify paraphrase (that is, non-literal translation) cases from NMT with the attention mechanism (Bahdanau et al., 2014). While most work would pay attention to the more strongly correlated parts in the resulting word alignments which often indicate very faithful and literal translation, we assume that the less correlated parts would correspond to free yet more fluent translation, provided that the bilingual parallel corpus is of good quality. Preliminary experiments are underway, and there are certainly technical issues to overcome, including threshold setting, noise filtering, and properly making use of the less strongly aligned parts. Evaluation would also need to be considered.

Acknowledgements

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CUHK 14616317).

References

J. Aitchison. 2003. *Words in the Mind: An Introduction to the Mental Lexicon*. Blackwell Publishers.

B.T.S. Atkins and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press.

D. Bahdanau, K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *arXiv e-prints, abs/1409.0473*.

V. Baisa, M. Jakubíček, A. Kilgarriff, V. Kovář, and P. Rychlý. 2014. Bilingual Word Sketches: the translate button. In *Proceedings of the 16th EURALEX International Congress*, pages 505–513, Bolzano, Italy.

C. Biemann, S. Bordag, and U. Quasthoff. 2004. Automatic acquisition of paradigmatic relations using iterated co-occurrences. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004)*, pages 967–970, Lisbon, Portugal.

L. Bowker. 2015. Computer-aided translation: Translator training. In S-W. Chan, editor, *The Routledge Encyclopedia of Translation Technology*. Routledge.

K.W. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29.

A.M. Collins and E.F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

F. Hill, R. Reichart, and A. Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

S. Jarvis and A. Pavlenko. 2008. *Crosslinguistic Influence in Language and Cognition*. Routledge, New York.

C. Keatley. 1992. History of bilingualism research in cognitive psychology. In R. Harris, editor, *Cognitive Processing in Bilinguals*, pages 15–49. North-Holland, Amsterdam.

A. Kilgarriff, P. Rychlý, P. Smrz, and D.Tugwell. 2004. The Sketch Engine. In *Proceedings of EURALEX 2004*, Lorient, France.

J. Kroll and G. Sunderman. 2003. Cognitive processes in second language learners and bilinguals: The development of lexical and conceptual representations. In C. Doughty and M. Long, editors, *The Handbook of Second Language Acquisition*, pages 104–129. Blackwell, Malden, MA.

O.Y. Kwong. 2013. Exploring the Chinese mental lexicon with word association norms. In *Proceedings of the 27th Pacific Asia Conference on Language, Information and Computation (PACLIC 27)*, pages 153–162, Taipei.

O.Y. Kwong. 2016. Strong associations can be weak: Some thoughts on cross-lingual word webs for translation. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation (PACLIC 30)*, pages 249–257, Seoul, Korea.

K.R. McKeown and D.R. Radev. 2000. Collocations. In R. Dale, H. Moisl, and H. Somers, editors, *A Handbook of Natural Language Processing*. Marcel Dekker.

H. Moss and L. Older. 1996. *Birkbeck Word Association Norms*. Psychology Press, Hove, UK.

R. Navigli and S. Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

D.L. Nelson, C.L. McEvoy, and T.A. Schreiber. 1998. *The University of South Florida word association, rhyme, and word fragment norms*. http://w3.usf.edu/FreeAssociation/.

M. Olohan. 2004. *Introducing Corpora in Translation Studies*. Routledge.

A. Pavlenko. 2009. Conceptual representation in the bilingual lexicon and second language vocabulary learning. In A. Pavlenko, editor, *The Bilingual Mental Lexicon: Interdisciplinary Approaches*, pages 125–160. Multilingual Matters, Bristol, UK.

P. Rychlý and A. Kilgarriff. 2007. An efficient algorithm for building a distributional thesaurus (and other Sketch Engine developments). In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 41–44, Czech Republic.

M. Wettler and R. Rapp. 1993. Computation of word associations based on the co-occurrences of words in large corpora. In *Proceedings of the 1st Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 84–93, Columbus, Ohio.

M. Zock, O. Ferret, and D. Schwab. 2010. Deliberate word access: an intuition, a roadmap and some preliminary empirical results. *International Journal of Speech Technology*, 13(4):201–218.