Translating Collocations: The Need for Task-driven Word Associations

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

Abstract

Existing dictionaries may help collocation translation by suggesting associated words in the form of collocations, thesaurus, and example sentences. We propose to enhance them with task-driven word associations, illustrating the need by a few scenarios and outlining a possible approach based on word embedding. An example is given, using pre-trained word embedding, while more extensive investigation with more refined methods and resources is underway.

1 Introduction

In practical bilingual lexicography, there is an important distinction between context-free and context-sensitive translation. Context-free translation refers to the general equivalents in a target language given for a particular headword in a source language; and context-sensitive translation refers to the rendition of a headword appropriately according to its occurrence in a given sentence or context. What lexicographers often do is to first produce many translations of a headword in context, and then distill from them a safest equivalent to be the "direct translation" of the headword in the entry, which could be suitably used in most contexts (Atkins and Rundell, 2008). This is in response to the habit of many users of bilingual dictionaries who will take the first equivalent found in the entry and use it without paying much attention to the actual context (Atkins and Varantola, 1997).

In actual translation, however, plugging in the first equivalent found in a dictionary regardless of the context in front of the translator is exactly what is most discouraged. Hence for a bilingual dictionary to be helpful to translators, adequate example sentences should be provided to enlighten users of different rendition possibilities and their appropriateness in a variety of contexts. On the other hand, for a translator to use a bilingual dictionary properly and smartly, one has to possess the skills to access the diverse contexts embedding a certain word and thus a whole range of context-sensitive equivalents in addition to the neutral but probably duller word choices.

In this study, we focus on the translation of ADJ-N collocations from English to Chinese, and consider the lexical information demand on the translator's part. In addition to the access means in existing dictionaries, in the form of collocations, thesaurus, and examples, we propose to enhance them with task-driven word associations filtered from pre-trained word embedding. This is expected to achieve three purposes: to extend the coverage of less common collocations, to assist translators in more precise word choices, and to encourage the use of different translation strategies appropriate in specific contexts.

2 Collocations and Translation Strategies

The translation of collocations has long been an issue (e.g. Chukwu, 1997; Shraideh and Mahadin, 2015), and different languages may not have the same collocations (McKeown and Radev, 2000). In the current discussion, we focus on the translation of English ADJ-N collocations to Chinese, which may seem straightforward at times but could always be challenging when considered from the context-sensitive side. Take a simple example like *good friend*. It can be directly and compositionally rendered as 好朋友 (*good*=好 + *friend*=朋友). While in most cases this would be perfectly fine and most acceptable, in

Proceedings of the Workshop on Cognitive Aspects of the Lexicon, pages 112–116
Barcelona, Spain (Online), December 12, 2020

translation teaching we are nevertheless told that there are always other alternatives which may fit the contexts even better, and there are different strategies to achieve equivalence at various levels (e.g. Baker, 2011). This is especially salient for literary translation. Hence, if the *good friend* in the source text refers to more or less a *confidant*, we may use another Chinese word 知己; or if the original emphasises the length and intensity of the friendship, phrases like 相知多年 (literally meaning "mutually know well for many years") and 友情深厚 (literally meaning "friendship is deep and thick") may be used, amongst many other possible renditions.

The above example shows that even for a simple ADJ+N collocation, a translator may need access to thesaural information, or near-synonyms, in both the source and target languages, to make appropriate lexical choices. In addition to paradigmatic associations, syntagmatic associations are also necessary, to find out what words can naturally describe good and long friendship in the target language. Moreover, even broader word associations are required to enable other translation strategies like transposition, modulation, and paraphrase, which involve the shift in word class and probably an extension into more culturally specific vocabulary items. These three scenarios are further illustrated below, with reference to the Macmillan Dictionary[1] and the Cambridge English-Chinese Dictionary[2].

2.1 Less Common Collocations

The thesaural and collocational information available in a dictionary often only covers the most typical cases. For instance, in the citations for honorary degrees or honorary fellowships in some universities in Hong Kong, the recipients are often praised for their *remarkable contributions*. Looking up both words in the Macmillan Dictionary, the combination is not found in the entries, and also not under Collocations and examples. The adjectives frequently used with *contribution* are: *great*, *huge*, *important*, *major*, *outstanding*, *positive*, *significant*, *useful*, and *valuable*. Nouns frequently used with *remarkable* are given in several groups, like something done or achieved: *achievement*, *career*, *feat*, *progress*, *recovery*, *success*; being similar: *resemblance*, *similarity*; person or people: *man*, *people*, *woman*; etc. Meanwhile, from the bilingual Cambridge Dictionary, the combination is not demonstrated in any of the examples, and the context-free equivalents for *remarkable* (非凡的; 奇異的; 引人注目的 – all leaning toward the sense of extraordinary and unusual) cannot naturally collocate with *contribution* (貢獻 – something you do to help achieve something, disambiguated from money you give and article you write).

2.2 Very Common Collocations

While one has to find ways to think of the appropriate renditions for less common collocations, very common collocations may also demand some creativity on the translator's side to go beyond the context-free combinations. For instance, when a high-level or general adjective (like *good*, *great*, or *nice*) is used to modify a noun, there could be a better and more specific adjective in the target language to go with the noun. In practice, anything can be good and the most general equivalent of *good* is 好. But to render *good idea* as 好主意 may not always be a good idea, depending on the actual context and style of the source text. Under the *idea* entry in the Cambridge Dictionary, two of the example sentences show the use of *good idea* but both are rendered as 好主意. There is another example with *bright (=good) idea*, translated as 好點子 (slightly informal). The problem is how we may inspire translators with the other alternatives.

2.3 Beyond Literal Translation

There are times when literal translation is not all acceptable from the target language side, and a translator must resort to other strategies that inevitably involve a shift in word class, or when there is a much more idiomatic expression, sometimes cultural specific, for the rendition. An example is *vivid memories*, as in "I still have vivid memories of my childhood". The bilingual Cambridge dictionary gives 栩栩如生的; 鮮活的; 生動的 (which are more like "seeing something brought to life") for *vivid* which are not likely to collocate with the equivalent of *memory* (記憶). So the best translation is not necessarily in the

[1] https://www.macmillandictionary.com/
[2] https://dictionary.cambridge.org/

form of ADJ+N, but rather done with a shift in word class, like 清楚記得 (clearly remember), 印象難忘 (impressive, unforgettable), as well as other four-character Chinese idioms like 記憶猶新 and 歷歷在目 which all suggest how clearly one remembers something.

3 Task-driven Word Associations

The issue here is therefore to expand lexical access routes in dictionaries, on top of the thesaural and collocational information as well as example sentences already found therein, to facilitate translators' work and to inspire them of the possibilities for rendition. Lexical access is largely concerned with word associations which form the basis of modelling the mental lexicon as a vast network (e.g. Aitchison, 2003). The interconnection of words in such a network can be used to account for and model various phenomena of the semantic memory like tip-of-the-tongue problem (e.g. Zock and Biemann, 2016). Free word associations include associative relations of different types and strengths. Their statistical modelling from large corpora (e.g. Church and Hanks, 1990; Wettler and Rapp, 1993) has contributed to lexicography for finding collocations and thesaural groups. There is a class of models and methods under distributional semantics (Harris, 1954; Baroni and Lenci, 2010; Clark, 2012), where word senses are represented by means of word co-occurrence vectors. The main assumption is that similar words appear in similar contexts, and by comparing the similarity of the vector spaces, it makes a popular approach for extracting paradigmatically related words (e.g. Agirre et al., 2001; Biemann et al., 2004; Hill et al., 2015; Santus et al., 2016). Word embedding (Mikolov et al., 2013) is a vector model among the latest trends.

3.1 Associations for Different Purposes

Kwong (2016) has shown from a comparison of English and Chinese free association norms that the association patterns are quite different. Free associations tend to be paradigmatic relations in English (e.g. *correct – right*), but syntagmatic or collocational relations in Chinese (e.g. 正確 *correct* – 答案 *answer*). Collocations and thesaural groups obtained from large corpora, like those computed by the Sketch Engine (Kilgarriff et al., 2004), may not always agree with the word association norms. Sometimes apparently strong associations may not rank high. The main problem, however, is that the associations are not task-specific.

As discussed in Section 2, when translating collocations, we need both paradigmatic and syntagmatic associations, and even broader relations. At the same time, the associations should not be free, because they should be relevant to the collocation being translated. Hence, while it is interesting to know what words are synonymous to *remarkable*, not all of them are relevant if they do not usually modify *contributions*. Similarly, it is useful to know what *remarkable* often modifies, but for this task they would not be informative if they are not also closely associated with *contributions*. Hence, we need to be able to refer to the associations relevant for a particular purpose. In other words, free associations should be re-prioritised for specific language tasks.

3.2 An Example

In this example, we try to address the kind of situations discussed in Section 2.1. Cosine similarities between words are computed with the pre-trained GloVe (6B tokens, 50d) word vectors (Pennington et al., 2014). As the models learn the word representations from their usual contexts, word embeddings are known for their good job on computing word similarity and analogies, which are surprisingly intuitive and interesting. But in contrast to what is usually highlighted, words with high similarities are not restricted to paradigmatically related words. As shown in Figure 1, although *remarkable* and *outstanding* are expected to be similar to each other, the similarity scores may actually be even higher between the adjectives and the nouns they modify. For instance, the similarities for *remarkable – achievements*, *outstanding – achievements*, and *outstanding – contributions* are higher than those for *remarkable – outstanding* and *contributions – achievements*, which are paradigmatically related. This observation implies two things: First, the words found similar to a given word may be considered free associations. Such associations may cover different kinds of relations. Second, ADJ+N may share high similarity if

they often co-occur in the same context. In this case, it is quite obvious that *remarkable contributions* may be a less seen combination compared with the others. Thus this is another piece of information revealed from the similarity scores with respect to the closeness of the syntagmatically related words. Also, the similarity scores should be considered in a relative sense. So depending on the task at hand, we should re-order the so-called similar words based on their parts of speech and relative scores to filter the useful information.

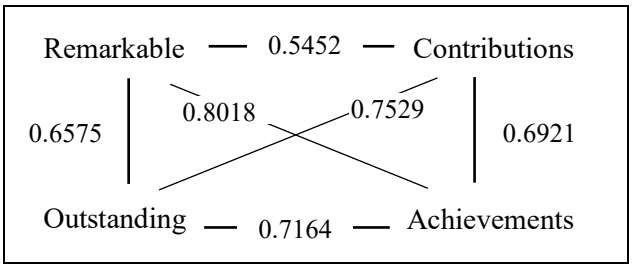

Figure 1: Similarity scores compared

Remarkable		Contribution	
astonishing		contributions	
accomplishment	✓	exceptional	✓
impressive		substantial	
incredible		outstanding	✓
amazing		achievement	✓
surprising		extraordinary	✓

Table 1: Some top associations

3.3 Proposed Steps

Hence, we should have a further interpretation of word embedding, and the information it may provide for our task. What we do here is not only to look for similar words for one word, but check out similar adjectives and nouns back and forth to gather similar collocations for reference, to supplement the less seen combinations not covered in dictionaries. Table 1 shows the top associations for *remarkable* and *contribution* based on similarity scores.

First, screen the adjectives similar to *remarkable* and select those which have higher similarity with *contribution* than *remarkable – contribution*. This gives us *outstanding*, *extraordinary* and *exceptional*. They are not shown in Table 1 because they were lower in the list, while it illustrates how we discard words highly similar to *remarkable*, like *astonishing* and *amazing*, for they are really less relevant with *contribution*.

Second, are there any nouns associated with *remarkable* that are also similar to *contribution*? In other words, find the words closely related to *contribution* that may be more commonly modified by *remarkable*. This gives us *accomplishment(s)*, *achievement(s)* and *success(es)*. Words like *milestone*, *inspiration* and *impression* are close to *remarkable*, but not to *contribution*, and they are pushed further down.

Seeded by the selected associations, the links to them in bilingual dictionaries can offer more navigation routes for users, not only to the context-free equivalents, but the corresponding example sentences which may showcase more context-sensitive translations. In this case, 卓越 (more used for *outstanding*) may be a better choice to go with 貢獻 (*contribution*), and of course, the latter may also have other synonymous alternatives in Chinese.

4 Ongoing Work

We have only outlined the steps to be tried, in their primitive and crude forms. But while the means are still under investigation, the ends are clear. The word associations modelled from large corpora may give nice results in certain cases, but in practice the associated words can often be so broadly related that they will simply not be equally activated when a person is performing a particular language task. Hence, we need task-driven associations to focus on the most relevant associated words in a given context. In this study, we have used collocation translation as a task and proposed to filter the associations by means of similarity or closeness obtained from word embedding. By doing so, we expect to enhance the lexical access means in dictionaries to assist translators in producing both faithful and fluent renditions.

In the present discussion, we only used a pre-trained embedding for some preliminary exploration for one scenario. More extensive work is underway, including the refinement of the method to handle different scenarios, the development of a systematic collocation set for testing, and the use of both English and Chinese word embedding in the process.

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

Agirre, E., Ansa, O., Martinez, D. and Hovy, E. 2001. Enriching WordNet concepts with topic signatures. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations*, Pittsburgh, pp.23-28.

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

Atkins, B. and K. Varantola. 1997. Monitoring Dictionary Use. *International Journal of Lexicography, 10*: 1-45.

Atkins, B.T.S. and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. New York: Oxford University Press.

Baker, Mona. 2011. *In Other Words: A Coursebook on Translation*. New York: Routledge.

Baroni, M. and A. Lenci. 2010. Distribution memory: A general framework for corpus-based semantics. *Computational Linguistics, 36(4)*: 673-721.

Biemann, C., Bordag, S. and Quasthoff, U. 2004. Automatic Acquisition of Paradigmatic Relations using Iterated Co-occurrences. In *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal, pp.967-970.

Chukwu, Uzoma. 1997. Collocations in translation: Personal textbases to the rescue of dictionaries. *ASp, 15-18*: 105-115.

Church, K.W. and P. Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics, 16(1)*:22-29.

Clark, S. 2012. Vector Space Models of Lexical Meaning. In S. Lappin and C. Fox (Eds.), *The Handbook of Contemporary Semantic Theory* (pp.493-522). Hoboken: John Wiley & Sons.

Harris, Z. 1954. Distributional structure. *Word, 10(2-3)*: 146-162.

Hill, F., Reichart, R. and Korhonen, A. 2015. SimLex-999: Evaluating Semantic Models with (Genuine) Similarity Estimation. *Computational Linguistics, 41(4)*: 665-695.

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

Kwong, O.Y. 2016. Strong Associations Can Be Weak: Some Thoughts on Cross-lingual Word Webs for Translation. To appear in *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation (PACLIC 30)*, Seoul, Korea.

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

Mikolov, T., Chen, K., Corrado, G. and Dean, J. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.

Santus, E. Lenci, A., Chiu, T-S., Lu, Q. and Huang, C-R. 2016. Nine Features in a Random Forest to Learn Taxonomical Semantic Relations. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, pp.4557-4564.

Shraideh, Khetam W. and Radwan S. Mahadin. 2015. Difficulties and Strategies in Translating Collocations in BBC Political Texts. *Arab World English Journal, 6(3)*: 320-356.

Wettler, M. 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*, Columbus, Ohio, pp.84-93.

Zock, Michael and Chris Biemann. 2016. Towards a resource based on users' knowledge to overcome the Tip-of-the-Tongue problem. In *Proceedings of the Workshop on Cognitive Aspects of the Lexicon*, Osaka, Japan, pp.57-68.