DOMA I N  WORD TRANSLATION BY SPACE-FREQUENCY ANALYSIS OF CONTEXT LENGTH HISTOGRAMS

Pascale Fung

Computer Science Department
Columbia University
New York, NY 10027
pascale@cs.columbia.edu

ABSTRACT
We report a new statistical feature relating a bilingual word pair in a non-parallel English-Chinese corpus. It is found that the lengths of context segments of a word is closely correlated to that of its translation, even when the corpus is non-parallel, i.e., monolingual texts which are not translations of each other. The context segment length histogram of a word has a characteristic pattern and corresponds to that of its translation. If a word appears most frequently in long segments, its translation is found to be most likely occurring in long segments. One way to match these histograms is to first extract their salient shape characteristics by space-frequency analysis and then match them against each other using dynamic time warping. The results of matching can be used in combination with other statistical features to bootstrap a word or term translation algorithm from non-parallel corpora.

1. INTRODUCTION
Translating domain-specific words is a significant component in machine translation and machine-aided translation systems. These words are often not found in standard dictionaries. Human translators, not being experts in every technical or regional domain, cannot produce their translations effectively. Automatic translation of words in specific domains is therefore highly desirable.

One approach to obtaining domain-specific word translations employs statistical learning algorithms to automatically extract a lexicon from large parallel bilingual texts which contain the same material in two translations [1, 3, 4, 5, 7]. The weakness of this approach is that parallel texts are relatively scarce. However, non-parallel texts can be found easily, since they are simply monolingual texts concerning the same domain but not necessarily translations. Another weakness of some previous approaches is their orientation toward European language pairs. They cannot be applied to language pairs such as Chinese and English. A new approach is which would be extendable to other language pairs is needed.

This paper demonstrates a pattern matching method by using a statistical feature, the context length histogram, to correlate pairs of translated words. It will also be shown how space-frequency analysis is used for matching such word pair signals for translation.

As input corpus, the bilingual transcription of the Hong Kong Legislative Council debates is used for experiments [6]. The data is from 1988–1992, with the first 73618 sentences from the English text, and the next 73618 sentences from the Chinese text. There are no overlapping sentences between the texts. The topics of these debates focus on the political and social issues of Hong Kong.

2. ALGORITHM OVERVIEW
The procedure for our algorithm is as follows:

1. Segment both the English and the Chinese texts by delimiters
2. Compute segment lengths of both texts
3. Compute the context length histograms for all words
4. Transform the histograms using space-frequency analysis
5. Dynamic time warping to match the transformed graphs
6. Obtain bilingual word pairs from matching results

2.1. Segments of text in English and Chinese
Segmental information was found to be useful in providing statistics for word pair matching. In parallel corpora, a long sentence in one language would correspond to a long sentence in its translation to another language. Such information could be used to align sentences and word pair matching could be carried out
from aligned sentence pairs [1, 4, 5, 7]. Texts in noisy parallel corpora could be segmented by word pair anchor points [3] and aligned. In a non-parallel bilingual text, there is no such linear sentence or segmental mapping—given any sentence in one language, its translation does not even appear in the other text. We need to find segmental correspondence which are text-independent.

It is generally found that English sentences, delimited by a full-stop(period), are shorter than Chinese sentences, delimited by a round circle. Very often, Chinese would use commas or semi-colons instead where in English a full-stop would have been used. Therefore, full-stops in English and Chinese are not good corresponding delimiters for segments. On the other hand, punctuations in general are still good delimiters. So we divided both the English and the Chinese texts into segments delimited by one of the following punctuations: an English full-stop, a Chinese full-stop, a comma, a question mark, a semi-colon or an exclamation mark.

We postulate that if a word appears frequently in short segments, then its translation would also appear more frequently in short segments. For example, the word figure is often seen in segments like “We will show this in figure 1”, “The ... is shown as follows in figure 1” etc. It rarely appears in long segments. Its translation is used in the same ways in Chinese. We define the length of an English segment to be the number of words in that segment. However, the length of a Chinese segment is defined as the number of characters in that segment to compensate for its linguistic difference with English.

3. HISTOGRAMS OF CONTEXT SEGMENT LENGTHS

Next, we compute the histogram of context segment lengths for each word in English and Chinese, assuming the maximum segment length is 100 and the minimum is one\(^1\). This is also the range for the x-axis for the histogram plot. For Government, part of its concordance in the English text is shown in Figure 1, one segment per line. The concordance for 政府, the Chinese word is shown in Figure 2. The first field indicates the length of each segment. The y value of the histogram indicates how many times Government occurs in a segment of length x. Since this information would be used to match words in non-parallel texts with very different occurrence frequencies, we normalize this graph so that the total area under the graph would be one. We obtain a graph of the same shape but with different y values.

Example plots of the words Government and debate in both languages are shown in Figures 4 and 6. Note the visual similarity between the two graphs of a word pair. By inspecting the original corpus, we found that the salient narrow peaks such as the one at x = 4 in the histogram for Government and those at x = 10 and then 16 for debate are caused by the predominance of domain-specific rigid phrases of those segment lengths.

---

\(^1\)In actual case, the maximum is usually around 70
4. SPACE-FREQUENCY TRANSFORMATION FOR MATCHING HISTOGRAMS

From the histogram figures, we can see that in general, the similarity between the histogram of a word in English and that of its translation in Chinese have similarities perceivable to the human eyes, i.e., we can see they have similar shapes. To match these shapes algorithmically, however, is much more difficult. Thus, we need a way to analyse the general hump, as well as the peaks and valleys of the plots, preserving the order in which they appear.

A space-frequency transformation can be used to analyze the signals in order to emphasize their characteristics and to reduce superfluous information. The difference of two Gaussians was used as a basis function (Figure 3):

\[
\begin{align*}
    h &= \left( \frac{1}{\sqrt{2\pi a^2}} e^{-0.5a^2/x^2} \right) - \left( \frac{1}{\sqrt{2\pi a^2}} e^{-0.5a^2/x^2} \right) \\
    a &= 1, 5, 10, \ldots, N \\
    a2 &= 0.5a \\
    u &= -5a : 5a
\end{align*}
\]

The total area of the basis function is zero. Different \(a\) would contract or dilate the basis function, thereby changing the window size of the transformation. The

\[
\text{val}(V1, y) = \int_0^N h \cdot V'
\]

Figure 3: Difference of two Gaussians as the basis function

basis functions was convolved with the interpolated graph of the original signal \(V\) at all positions on the \(x\)-axis

After transformation, the value \(i\) at \((x, y)\) would denote the intensity at frequency \(y\) at point \(x\). At any given \(x\), the plot is a weighted combination of different space-frequency basis functions at frequencies marked by the \(y\)-axis. The weight is shown by \(i\). The space-frequency transformation provides us an analytical way of looking at the histogram signals. By quantizing the signals, the relative peaks and valleys on the signals become more salient.

Looking at the left plot in Figure 5, the transformation plot of the word Government in English, we see that there is a small white patch at around \(y = 5\), high frequency. This corresponds to the sharp peak, a local maxima in the original histogram at around \(x = 5\). The bigger white patch at lower frequencies and at around \(x = 28\) corresponds to the general shape of the original signal having a gentle hump there. There are corresponding sharp peaks and a gentle hump in the signals for the Chinese word, and in its transformation figure.

5. DTW MATCHING

After transformation, the signals of word pairs are more or less warped versions of each other in the \(x\)-axis. To match the transformed graphs, we use dynamic time warping (DTW) on the difference of the intensity at each frequency. At each \(y\), the \((x-1)\)-dimensional row vector is the delta encoder of the original \(x\)-dimensional vector. We compare the row vector of a word \(V1\) to that of another word \(V2\) at the same \(y\) value, giving a score \(DTW(V1, V2, y)\). The total correlation score between two graphs is \(\sum_{i=1}^N DTW(V1, V2, y)\) where \(DTW(V1, V2, y)\) is the DTW score of the two delta vectors in frequency band \(i\).

6. RESULT AND DISCUSSION

We have shown a novel algorithm for extracting bilingual word pairs from same domain, non-parallel texts of Chinese and English. The signal representation of word features ensures that this algorithm is robust to language groups. DTW is an effective matching function on the space-frequency transformations of the word signals. We have tested this algorithm on more than 50 word pairs, the result shows that about 40 of the words match most closely to their translations in the other language. We will combine this feature with other statistical information found in our previous work [2] in order to improve the performance.

7. ACKNOWLEDGMENT

I wish to thank Truong-Thao Nguyen and other faculties at the Dept. of Electrical & Electronic Engineering in HKUST for useful comments, Kathleen McKeown, Shreer Nayar, and Hiroshi Murase for support and encouragement.
8. REFERENCES


![Figure 4: Normalized histogram of Government in English and Chinese](image1)

![Figure 5: Space-frequency plots of Government in English and Chinese](image2)

![Figure 6: Normalized histogram of debate in English and Chinese](image3)

![Figure 7: Space-frequency plots of debate in English and Chinese](image4)