An English-Translated Parallel Corpus for the CJK Wikipedia Collections

Ling-Xiang Tang
Queensland University of Technology
Brisbane, Australia
l4.tang@qut.edu.au

Shlomo Geva
Queensland University of Technology
Brisbane, Australia
s.geva@qut.edu.au

Andrew Trotman
University of Otago
Dunedin, New Zealand
andrew@cs.otago.ac.nz

ABSTRACT
In this paper, we describe a machine-translated parallel English corpus for the NTCIR Chinese, Japanese and Korean (CJK) Wikipedia collections. This document collection is named *CJK2E Wikipedia XML corpus*. The corpus could be used by the information retrieval research community and knowledge sharing in Wikipedia in many ways; for example, this corpus could be used for experimentations in cross-lingual information retrieval, cross-lingual link discovery, or omni-lingual information retrieval research. Furthermore, the translated CJK articles could be used to further expand the current coverage of the English Wikipedia.

Categories and Subject Descriptors
H.3.7 [**Information Storage and Retrieval**]: Digital Libraries – *collection.*

General Terms
Documentation, Experimentation, Languages

Keywords
Wikipedia, Corpus, English, Chinese, Japanese, Korean, cross-lingual information retrieval, cross-lingual link discovery, machine learning.

1. INTRODUCTION
Wikipedia is currently the largest freely available online multilingual encyclopaedia. It contains a large number of articles covering millions of topics and has articles in most written languages. However the different language versions of Wikipedia have evolved at different rates and are unbalanced in coverage (and sometimes differently biased in content). Among all the language versions, English Wikipedia is the largest with over 6,550,000 articles[1].

But an article may not be written in a user's preferred language, or the user may be looking for richer content than is available in their preferred language. In these cases our user may be able to, and prepared to, read in a second or subsequent language – if they could find the content.

To address the problem of finding content in multiple languages NTCIR launched CrossLink, the cross-lingual link discovery (CLLD) [1] track. The aim of this track at NTCIR-9 was to build a system that could automatically recommend hypertext links from English documents to relevant documents in Chinese, Japanese, and Korean. Such a system must not only recommend topically relevant documents to a source document, but must also suggest appropriate anchors.

Good approaches to CLLD were seen at NTCIR-9 Crosslink. We observed that most systems seen there used translation in some ways. Typically seen were: direct machine translation (of, for example, entities) or triangulation (for example, following links from an English source article to an English target article then finding the Chinese equivalent article thorough "language links" or entity translation) [1-6].

To lower the barrier of entry to research such as CLLD and other cross lingual Information Retrieval problems we created and present here a machine-translated parallel English Wikipedia corpus derived from the Chinese, Japanese, and Korean (CJK) Wikipedia collections currently being used at NTCIR-10 CrossLink-2. This new corpus was built by translating the CJK Wikipedia articles into English using an online machine translation service (specifically, Google Translate[2]).

This machine translated corpus could be used for many purposes, including (but not limited to): cross-lingual link discovery (CLLD); cross-lingual information retrieval; omni-lingual information retrieval; as well as machine learning; cross-lingual document categorisation and clustering; and machine translation itself.

Better, the translated CJK articles could be used to further expand the coverage of the English Wikipedia. Topics not covered in English could be added and topics inadequately covered in English could be expanded (albeit not automatically).

The remainder of this article describes our translated corpus *CJK2E Wikipedia XML corpus* (Version 1.0). Collection statistics, corpus creation procedures, and an experiment demonstrating its utility, are included.

[1]The article number was collected from the Wikipedia database dump taken on 4th January 2012.

ADCS '12, December 05 - 06 2012, Dunedin, New Zealand

[2] http://research.google.com/university/translate/index.html

2. THE CJK (SOURCE) COLLECTIONS

2.1 Corpora Statistics

The CJK Wikipedia collections[3] we used as the source for translation are those used in the NTCIR-10 Crosslink-2 track[4]. The collections were created from the Wikipedia XML dumps taken in January 2012. The original article text with Wikipedia mark-up was converted to XML using the YAWN system [7]. The details of the source collections are given in Table 1. The first column lists the language, the second column lists the number of documents in the dump, the third gives the size of the collection and the fourth column lists the dump date. For example, the Chinese dump taken on January 11th 2012 contains 432,988 documents and is 3.6 Gigabytes in size.

Table 1. Characteristics of the CJK Wikipedia collections used at NTCIR-10 Crosslink-2 task and for translation.

Language	Documents	Size	Dump Date
Chinese	404,620	3.6GB	11/01/2012
Japanese	858,610	9.8GB	04/01/2012
Korean	297,913	2.2GB	22/01/2012

2.2 Document Structure

Tags already present in the original Wikipedia XML dump files were maintained, but YAWN added new tags for article categories, sections, paragraphs, and links (amongst others). These new tags were added in an effort to provide additional structural information to the corpus user. The process followed has previously been successful at INEX. Examples of new and original tags, along with a brief description, are given in Table 2. The first column lists the tag, the second lists the source, and the third gives a brief description. For example, the title tag is original to the Wikipedia and gives the title of the article.

Table 2. Example tags from the YAWN version of the Wikipedia dumps, the second column lists the source of the tag: W for Wikipedia and Y for YAWN

Tag	Source	Description
Title	W	Document title
Id	W	The document identifier
Link	Y	Used for hypertext links. Cross-language links contain an attribute (e.g. "xlink:label="ko"") giving the target language (one of: zh, ja, ko, en
timestamp	W	Last update timestamp
categories	Y	A list of categories
category	Y	An individual category (seen within the categories tag)
P	Y	Paragraph
Sec	Y	Article section

[3]http://warehouse.ntcir.nii.ac.jp/openaccess/crosslink/10crosslink_documents.html

[4]http://ntcir.nii.ac.jp/CrossLink-2/

An excerpt of YAWN processed article taken from the Chinese Wikipedia is show in Table 6. The article, 裹蒸（a special type of Zongzi)[5] is showed as XML in the second row and as it appears in the Wikipedia in the fourth row. The other two rows are titles.

From the XML, it can be seen that the article is extensively marked up in XML and includes tags for such elements as: title, categories, and paragraphs. It also contains links to other Wikipedia articles.

3. CONVERSION TO ENGLISH

3.1 Considerations

Statistical machine translation systems are more effective when provided with context. That is, if asked to translate the contents of a short phrase, the accuracy can reasonably be expected to be lower than if given the entire sentence. Additionally, if the article was broken into its individual XML elements and each was translated separately it can reasonably be expected to take longer (using the Google API) than it would take for a single translation of the entire article.

However, the structural information has proven useful for both the presentation of articles and for providing *hint*s for the document retrieval. Wholesale removal of tags would detract from the utility of the translated collection and so as many tags as possible should be preserved through the translation. The preservation of tags also aids in the ability to map between the original and translated articles.

For the purpose of translating the collections herein, text formatting tags (b, i, li, etc.) and link tags (link, etc.) were removed before translation.

Table 3. Sections removed from the Chinese articles before translation. No sections were removed from the Japanese or Korean articles

Section	Chinese
Notes	注释 註釋
References	參考資料 参考资料 參考文獻 参考文献
External Links	外部連結 外部链接

To further increase throughput some sections were removed from the Chinese articles before translations. These sections included: “External Links”, “Notes”, and “References”. In each case the translation was likely to be of low accuracy and of little utility to the end user of the translated corpus.

[5]http://en.wikipedia.org/wiki/Zongzi

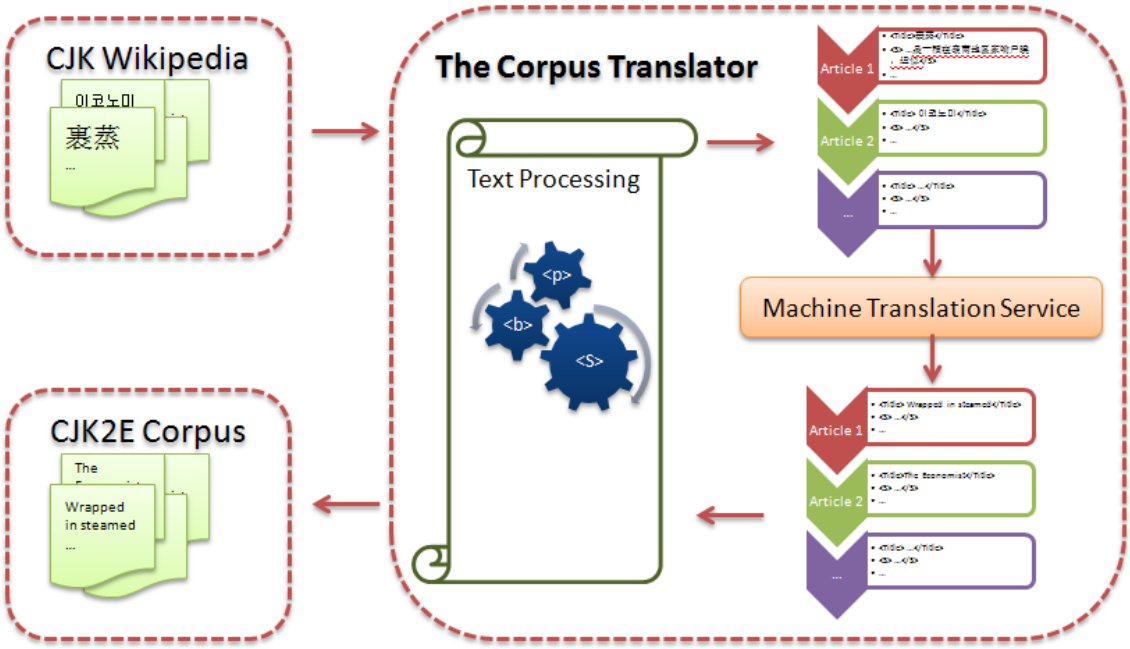

Figure 1. System design of the corpus translator

Table 3 presents a list of sections in column one and the Chinese equivalent in column two (including simplified, traditional, and other Chinese variants) removed from the documents before translation. For example, 注释 and 註釋 sections were removed as they were notes.

Such sections were only removed from the Chinese collections and not the Japanese or Korean collections because we did not have the necessary expertise to reliably identify such sections in those collections.

3.2 System Design

A design diagram of the corpus translator is given in Figure 1. The source article (in Chinese, Japanese, or Korean) is stripped of the formatting tags, and then decomposed into the remaining tags. They are further broken into sentences, which are segmented at the CJK punctuations where the segment size is not longer than 400 bytes (due to constraints imposed the translation service). Each chunk is translated and the translated chunks are re-composed into English articles and mark-up reintroduced.

3.3 The Statistics of the Translation

The translated corpus[6] contains 12,726,520 translated (to English) articles. Table 4 presents some statistics of the translated corpus. The first column lists the source language, the second gives the number of documents, and the third lists the size of the collection. For example, there are 397,571 articles translated from Chinese which take 2.6GB to store.

Not all articles successfully translated – but this is not unexpected. The translated corpus contains over twelve million articles that have been through YAWN conversion from Wikipedia formats into XML, processed by the corpus translator, translated by a third party, and pieced back together. Any one step in the process could reject a badly formed article or text segment. Version 1.0 of the collection consequently contains 98% of the original Chinese articles, less than 81% of the original Korean articles and only about 76% of the original Japanese articles. We are investigating the causes of the failures and will release updated collections in the future.

Table 4. Characteristics of the translated collections

Language	Documents	Size
Chinese	397,571	2.6GB
Japanese	652,902	5.0GB
Korean	239,285	1.3GB

3.4 Translation Results

Table 7 shows a side by side comparison of the Chinese text and the translation results for the 裹蒸 article. The first column gives the name of the element, the second gives the original Chinese text and the third gives the English translation. For example, the title, 裹蒸, is translated as “Wrapped in steamed”.

It can be seen from the table (and from visual inspection of other translations) that short entity-like elements (title, category, etc.) have a high translation quality. For longer running text (paragraphs, etc.) the sentences contains many grammatical errors and sometimes make little sense. However, after reading the translation it is usually possible to understand the article (albeit with effort).

[6] http://www.clld.sef.qut.edu.au/corpus (to appear)

4. THE UESFULNESS OF THE CORPUS

The readability of the translated articles may be relatively poor because they were machine-translated and the machine translated article quality is not comparable to that of professionally translated article. Although the corpus is not created for human consumption, it may be still be useful despite the quality of machine translation. To demonstrate the usefulness of the CJK2E corpus, we designed a small experiment, which made use of this corpus for a cross-lingual link discovery (CLLD) task. The experiment compared two CLLD systems: the baseline system discovered links by searching the translated anchors in the target document collection; the other system found links through searching the anchors directly in the machine translated CJK2E corpus.

4.1 The Experiment

4.1.1 Experiment Overview

In NTCIR-9 CrossLink Task, there were 25 orphaned English Wikipedia articles used as test topics, and the metrics LMAP, R-Prec, and P@N were utilised for the CLLD system evaluation [1, 2]. A CLLD System is required to recommend prospective meaningful anchors for the test topics and for each anchor identify up to 5 relevant links in a different language. For each topic, up to 250 anchors are allowed.

Our experiment focused on the English-to-Chinese subtask. The document retrieval system employed for link retrieval was the ATIRE[7] open source search engine with its modified BM25 ranking function.

4.1.2 Baseline Run

Team QUT in the NTCIR-9 CrossLink task (English-to-Chinese) submitted a run (LinkProbIR_ZH) which first used a link mining method to recommend anchors in source documents, then translated those anchors into Chinese, and finally, searched the Chinese Wikipedia collection using the translated anchors as query terms for relevant links [5].

Their link discovery process is similar to a common approach to achieving cross-lingual information retrieval where queries are translated into the target language and then a monolingual IR system is used to locate the relevant documents in the target collection. This run will be used as the baseline run for system performance comparison.

4.1.3 The Run with the CJK2E Corpus

To compare the performance of the above system (that relies on translated anchors), an alternative run, LinkProbIR_CJK2E, was created by directly searching the anchor candidates in the CJK2E corpus. The number of identified links for each identified anchor was limited to 5, in accordance with the NTCIR-9 CrossLink task specification.

Trying to ensure that the two systems were comparable, the anchor candidates used to find links for the test topics in this run was the same as that used by the system to create run LinkProbIR_ZH. An anchor candidate (any given phrase), a, is selected by measuring its anchor weight, γ (>0) , [8] the probability of being an anchor. It is defined as:

$$\gamma = \frac{number\ of\ articles\ having\ (\boldsymbol{a})\ being\ used\ as\ an\ anchor}{number\ of\ articles\ that\ have\ text\ of\ anchor(a)} \quad (1)$$

[7] http://www.atire.org/

The anchor weight of anchor candidates was calculated using the data mined from the English Wikipedia corpus used in the Link-the-Wiki track of INEX [9, 10]. For each orphaned topic article, all possible n-gram substrings from the document were first computed. For each of these the γ score was looked-up, and the list of anchor candidates was then sorted by their anchor weight values.

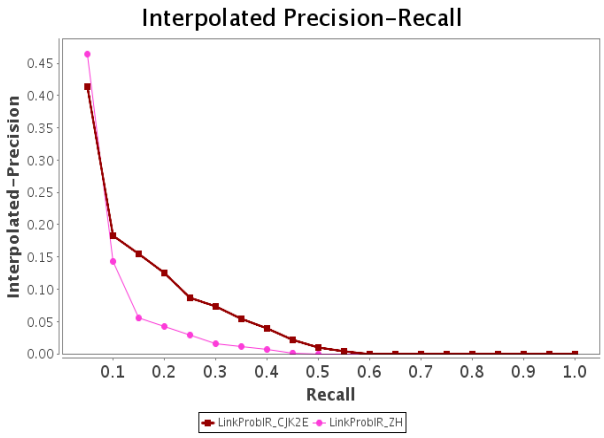

Figure 2. The interpolated P-R curves of two systems

4.2 Results and Discussions

The scores of the two experimental runs computed using the evaluation tool with the official *qrel* (Wikipedia ground-truth in file-to-file level) are given in Table 5. The runs are sorted on LMAP. For easy performance comparison, the interpolated precision and recall curves of two runs are also given in Figure 2.

From Figure 2, it can be seen that run LinkProbIR_CJK2E performs much better than run LinkProbIR_ZH. The scores computed in different metrics showed in Table 5 also suggest better links were discovered by run LinkProbIR_CJK2E than run LinkProbIR_ZH. Although, run LinkProbIR_ZH out-performs run LinkProbIR_CJK2E if measured with metrics P@5 and P@20, run LinkProbIR_CJK2E picked up many more good links if measured against a larger set of recommend anchors (P@30 or P@50, for example). A statistical analysis (two-tail paired *t*-test) on the LMAP scores of two runs over the 25 topics indicates that run LinkProbIR_CJK2E found significantly ($p = 0.01$) more links than LinkProbIR_ZH.

The experiment results indicate that the CJK2E corpus is useful in helping improve cross-lingual link discovery performance when cross-lingual information retrieval methods are involved. The performance difference of two systems could be attributed to the difference in translation quality. An article can provide more information to the translation engine than a single query with a few terms. The translation is expected to be superior as more contextual information is given. Translating a passage (as Google Translate does) is more likely to be correct than translating a word or a phrase at a time without the context of the embedding passage.

For CLLD approaches that first machine-translate anchor candidates into a target language and then searches them in the target document collection may discover fewer relevant links than other methods due to the possible inaccurate translation of *anchors* (caused by, for example, out of vocabulary (OOV)

terms). Although there are translation errors in the CJK2E corpus, overall most terms and phrases appear to have been correctly translated, resulting in an increased chance of hitting a relevant document if the anchors remain untranslated.

5. CONCLUSIONS

This article presents a machine-translated parallel English corpus which can be used by various cross-lingual link discovery, cross-lingual information retrieval, and machine learning systems to further improve their performance to satisfy users' information needs. An experiment was designed to justify the usefulness of the corpus, and the experimental results proved the claim. Wikipedia users may also find it useful because lots of articles can be adopted into the existing English Wikipedia with further proper editing and quality improvement.

6. ACKNOWLEDGMENTS

The Translate Research API is provided by Google.

7. REFERENCES

1. Tang, L.-X., Geva, S., Trotman, A., Xu, Y., Itakura, K.Y.: Overview of the NTCIR-9 Crosslink Task: Cross-lingual Link Discovery. In: Proceedings of NTCIR-9, pp. 437-463. (2011)

2. Tang, L.-X., Itakura, K.Y., Geva, S., Trotman, A., Xu, Y.: The Effectiveness of Cross-lingual Link Discovery. In: Proceedings of The Fourth International Workshop on Evaluating Information Access (EVIA), pp. 1-8. (2011)

3. Knoth, P., Zilka, L., Zdrahal, Z.: KMI, The Open University at NTCIR-9 CrossLink. In: Proceedings of NTCIR-9, pp. 495-502. (2011)

4. Kim, J., Gurevych, I.: UKP at CrossLink: Anchor Text Translation for Cross-lingual Link Discovery. In: Proceedings of NTCIR-9, pp. 487-494. (2011)

5. Tang, L.-X., Cavanagh, D., Trotman, A., Geva, S., Xu, Y., Sitbon, L.: Automated Cross-lingual Link Discovery in Wikipedia. In: Proceedings of NTCIR-9, pp. 512-519. (2011)

6. Fahrni, A., Nastase, V., Strube, M.: HITS' Graph-based System at the NTCIR-9 Cross-lingual Link Discovery Task. In: Proceedings of NTCIR-9, pp. 473-480. (2011)

7. Schenkel, R., Suchanek, F., Kasneci, G.: YAWN: A Semantically Annotated Wikipedia XML Corpus. In: 12. GI-Fachtagung für Datenbanksysteme in Business, Technologie und Web (BTW 2007). (2007)

8. Itakura, K., Clarke, C.: University of Waterloo at INEX2007: Adhoc and Link-the-Wiki Tracks. In: Proceedings of INEX 2007, pp. 417-425. (2008)

9. Huang, W., Geva, S., Trotman, A.: Overview of the INEX 2009 Link the Wiki Track. In: Proceedings of INEX 2009, pp. 312-323. Springer Berlin / Heidelberg, (2010)

10. Trotman, A., Alexander, D., Geva, S.: Overview of the INEX 2010 Link the Wiki Track In: Proceedings of INEX 2010, pp. 241-249. Springer Berlin / Heidelberg, (2011)

Table 5. The F2F automatic evaluation scores of two experimental runs with metrics (LMAP, R-Prec, and P@N)

Run ID	LMAP	R-Prec	P5	P10	P20	P30	P50	P250
LinkProbIR_CJK2E	0.044171	0.119954	0.128	0.152	0.138	0.141333	0.136	0.07536
LinkProbIR_ZH	0.02338	0.067135	0.184	0.16	0.118	0.109333	0.084	0.04352

Table 6. Excerpt from the Chinese Wikipedia article 裹蒸 about a special kind of Zongzi. The top shows the YAWN XML output and the bottom shows the original text as it appears in the Wikipedia

The article XML file

...

<title>裹蒸</title>

<categories>

<category>:點心</category>

<category>:糯米食品</category>

</categories>

<bdy>

裹蒸，俗作裹蒸粽，為<link xlink:type="simple" xlink:href="../../pages/263/263.xml">

中國</link><link xlink:type="simple" xlink:href="../../pages/412/412.xml">

廣東省</link><link xlink:type="simple" xlink:href="../../pages/813/15813.xml">

肇慶市</link>的特產，是一種在嶺南地區家喻戶曉，近似<link xlink:type="simple" xlink:href="../../pages/429/40429.xml">

粽</link>的煮食方式。傳統的裹蒸粽用料是<link xlink:type="simple" xlink:href="../../pages/266/39266.xml">

糯米</link>、<link xlink:type="simple" xlink:href="../../pages/880/372880.xml">

綠豆</link>、半肥瘦<link xlink:type="simple" xlink:href="../../pages/024/111024.xml">

豬肉</link>作餡，必須用西江两岸特有的<link>

冬葉</link>来包裹蒸制，具有独特的清香与优良的防腐作用。現在的裹蒸粽餡料包括五香肥肉、<link xlink:type="simple" xlink:href="../../pages/196/300196.xml">

鹹蛋</link>黃、燒<link xlink:type="simple" xlink:href="../../pages/554/15554.xml">

雞</link>、<link xlink:type="simple" xlink:href="../../pages/589/277589.xml">

燒鴨</link>、<link xlink:type="simple" xlink:href="../../pages/744/100744.xml">

叉燒</link>等。<p>

...

The real page on Chinese Wikipedia site

裹蒸 [编辑]

维基百科，自由的百科全书

（重定向自裹蒸粽）

裹蒸，俗作**裹蒸粽**，為中國廣東省肇慶市的特產，是一種在嶺南地區家喻戶曉，近似粽的煮食方式。傳統的裹蒸粽用料是糯米、綠豆、半肥瘦豬肉作餡，必須用西江两岸特有的冬葉来包裹蒸制，具有独特的清香与优良的防腐作用。現在的裹蒸粽餡料包括五香肥肉、鹹蛋黃、燒雞、燒鴨、叉燒等。

南齊書有記載，在皇帝御食中，有"裹蒸"這一道美食，皇帝稱，「我食此不盡，可四片破之，剩餘權當晚食」，裹蒸

Table 7. A side by side comparison of the original Chinese text and the translation for the article 裹蒸（"wrapped in steamed")

Element	Chinese Source Text	English Translated Text
title	裹蒸	Wrapped in steamed
category	點心	Dim Sum
category	糯米食品	Glutinous rice food
abstract	裹蒸，俗作裹蒸粽，為中國廣東省肇慶市的特產，是一種在嶺南地區家喻戶曉，近似粽的煮食方式。傳統的裹蒸粽用料是糯米、綠豆、半肥瘦豬肉作餡，必须用西江两岸特有的冬葉来包裹蒸制，具有独特的清香与优良的防腐作用。現在的裹蒸粽餡料包括五香肥肉、鹹蛋黃、燒雞、燒鴨、叉燒等。	Wrapped in steamed, popular for Guo Zhengzong, Guangdong, China Zhaoqing City's specialty, is a Lingnan region well-known approximate dumplings of cooking way. Traditional Guo Zhengzong materials is glutinous rice, mung bean, half fat and lean pork for filling, must use West River cross-strait unique winter leaves to parcel steaming, with a unique fragrance and excellent anti-corrosion role now Guo Zhengzong filling spiced fat salted egg yolk, burning chicken.Roast duck, barbecued pork.
p	南齊書有記載：在皇帝御食中，有“裹蒸”這一道美食。皇帝稱：「我食此不盡，可四片破之，剩餘權當晚食。」裹蒸是當地人作為春節、端午節探親拜年，必備的禮品，寓意豐衣足食和來年好運。清代詩人王仕禎有詩贊肇慶城鄉除夕熬煮裹蒸的盛況：「除夕濃煙籠紫陌，家家塵甑裹蒸香。」如今在肇慶的大街小巷中，觸目皆為肇慶裹蒸，與端硯齊名。不過，到現在的裹蒸依然採用人手製作。	Nan Qi Shu have recorded: the emperor imperial food, there are "wrapped in steamed" This Road food. Emperor said: "I eat this endless, four destructible, the remaining right to the night food. Wrapped in steamed locals as the Spring Festival, Dragon Boat Festival to visit relatives New Year, essential gift, symbolize affluence and the coming year, good luck. Qing Dynasty poet Raymond Wong Zhen poetry praise Zhaoqing urban and rural New Year's Eve boiled wrapped in steamed the grand occasion:" New Year's Eve smoke cage purple street, every household dust steamer wrapped steamed fragrant. "Now in theThe streets in Zhaoqing, Zhaoqing wrapped in steamed caught the attention of all and Duan par. However, now wrapped in steamed still handmade.
p	肇庆人制作裹蒸粽的传统从秦代就有了，其起源流传着两种说法，一种是秦军征统岭南时期，为驻军而设的可随身携带食用的煮熟的叶裹米团，一种是当时的农民在田间做活时，随身携带煮熟了的叶裹米团以便随时食用，但无论如何，裹蒸的来源并不是一般认为纪念屈原的端午粽，在肇庆民间一直是春节过年的食品，这就是裹蒸和粽子的区别。	The Zhaoqing person production Guo Zhengzong the tradition From the Qin Dynasty, there had its origin spread with two arguments, one kind is the period of Qin Zheng unified Lingnan for the garrison and set the can carry carry consumption of cooked leaves wrapped m group, one kind is time farmers in the field QUICKER player to carry cooked the leaves wrapped m group so that at any time eating, and in any case, but, wrapped in steamed the source is not generally think to commemorate Qu Yuan Dragon Boat Rice Dumplings Zhaoqing Folk Spring Festival, Chinese New Year's food, wrapped in steamed and dumplings the difference.
section title	著名品牌	Well-known brand
p	肥仔偉：一個旅行團到訪當地經常都會拜訪的裹蒸生產商。	Fatty Wei: a tour group visited the local often visit wrapped in steamed manufacturer.