Preprocessing XBRL files

The purpose of preprocessing XBRL files is to ensure that clean text is sent to the machine translation engine. This helps with producing high quality machine translations.

When identifying editing points, it's important to keep in mind that the XBRL file will first be uploaded to CAT software, go through the designated segmentation process (splitting the document into individual sentences), then sent to the machine translation API. We will mostly be focused on preprocessing XBRL files before they are uploaded to CAT software as the files are often messy. Although there are also segmentation issues, these can get pretty complex. For now, just read about the Pragmatic Segmenter to get an idea of how segmentation works.

After a document goes through the segmentation process, the CAT software creates tags to note any changes in font, and the tags occasionally confuse the machine translation engine.

Font changes are particularly frequent in Japanese documents as each Japanese character requires two spaces (full-width), while each English character requires only one space (half-width). Document creators will often mix the full and half-width characters such as when referring to a foreign product name in English, using an acronym, or in some cases writing numbers with the keyboard input set to English (half-width). The issue is exacerbated as the Japanese character format includes the option to type full-width alphabet as well.

This will be a live document where I take note of common errors, their location, and characteristics. I will broadly categorize them into two categories: Easy edits and Hard edits. Easy edits will be quick-to-implement, “find and replace” level edits. Hard edits will probably require some level of programming.

Character format examples

Vowels

Japanese (full-width): あいうえお
English (half-width): aiueo

Acronyms

Japanese (full-width): ＳＡＰ
English (half-width): SAP

Numbers

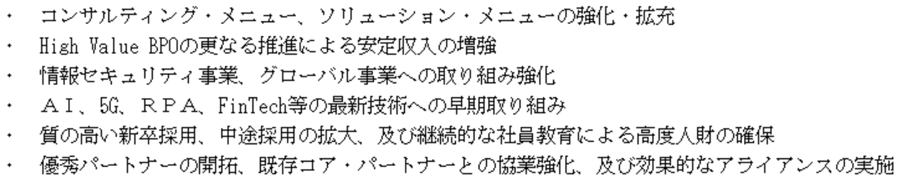
Japanese (full-width): １２３

English (half-width): 123

Machine translation output examples

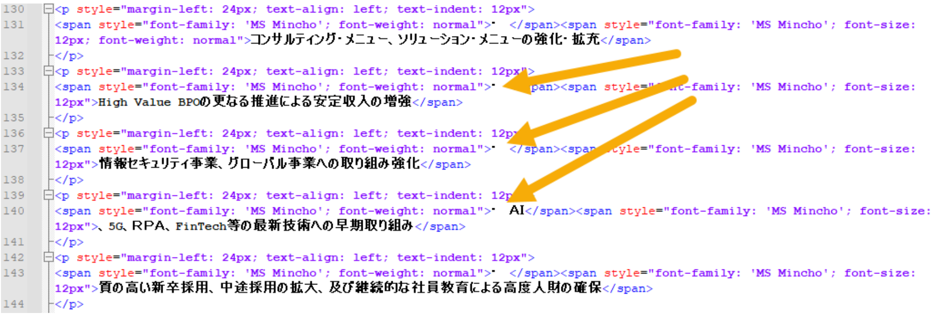
1. Full-width bullet point vs. half-width dash

The following section found in Business Brain Showa-Ota's FY03/20 Annual Securities Report contains full-width bullet points:



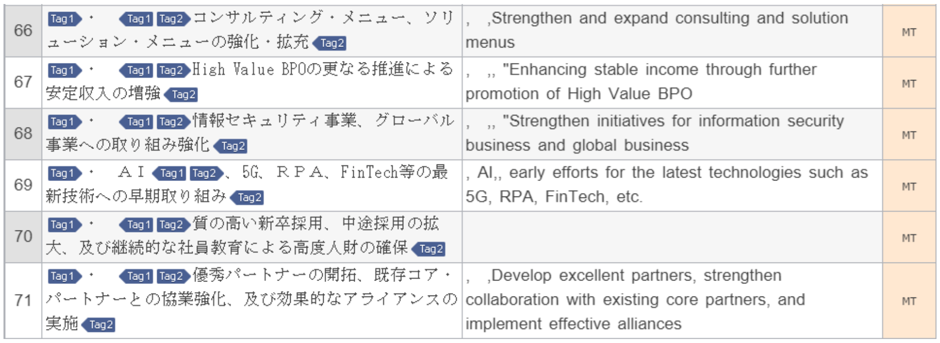
- コンサルティング・メニュー、ソリューション・メニューの強化・拡充
- High Value BPOの更なる推進による安定収入の増強
- 情報セキュリティ事業、グローバル事業への取り組み強化
- ＡＩ、5G、ＲＰＡ、FinTech等の最新技術への早期取り組み
- 質の高い新卒採用、中途採用の拡大、及び継続的な社員教育による高度人財の確保
- 優秀パートナーの開拓、既存コア・パートナーとの協業強化、及び効果的なアライアンスの実施

In an HTML editor (Notepad++), the same section appears as below. The orange arrows point to a few of the full-width bullet points:



```
<p style="margin-left: 24px; text-align: left; text-indent: 12px">
<span style="font-family: 'MS Mincho'; font-weight: normal">・ </span><span style="font-family: 'MS Mincho'; font-size:
12px; font-weight: normal">コンサルティング・メニュー、ソリューション・メニューの強化・拡充</span>
</p>
<p style="margin-left: 24px; text-align: left; text-indent: 12px">
<span style="font-family: 'MS Mincho'; font-weight: normal">・ </span><span style="font-family: 'MS Mincho'; font-size:
12px">High Value BPOの更なる推進による安定収入の増強</span>
</p>
<p style="margin-left: 24px; text-align: left; text-indent: 12px">
<span style="font-family: 'MS Mincho'; font-weight: normal">・ </span><span style="font-family: 'MS Mincho'; font-size:
12px">情報セキュリティ事業、グローバル事業への取り組み強化</span>
</p>
<p style="margin-left: 24px; text-align: left; text-indent: 12px">
<span style="font-family: 'MS Mincho'; font-weight: normal">・ AI</span><span style="font-family: 'MS Mincho'; font-size:
12px">、5G、RPA、FinTech等の最新技術への早期取り組み</span>
</p>
<p style="margin-left: 24px; text-align: left; text-indent: 12px">
<span style="font-family: 'MS Mincho'; font-weight: normal">・ </span><span style="font-family: 'MS Mincho'; font-size:
12px">質の高い新卒採用、中途採用の拡大、及び継続的な社員教育による高度人財の確保</span>
</p>
```

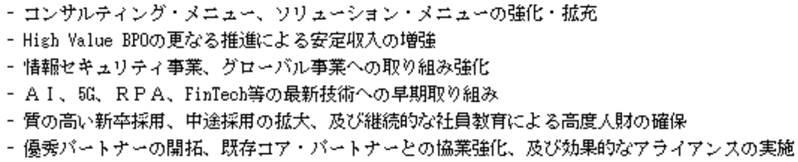
Here's a screenshot from Wordfast Pro 6 after DeepL machine translations have been populated:



66	Tag1 ・ Tag1 Tag2 コンサルティング・メニュー、ソリューション・メニューの強化・拡充 Tag2	, ,Strengthen and expand consulting and solution menus	MT
67	Tag1 ・ Tag1 Tag2 High Value BPOの更なる推進による安定収入の増強 Tag2	, ,, "Enhancing stable income through further promotion of High Value BPO	MT
68	Tag1 ・ Tag1 Tag2 情報セキュリティ事業、グローバル事業への取り組み強化 Tag2	, ,, "Strengthen initiatives for information security business and global business	MT
69	Tag1 ・ ＡＩ Tag1 Tag2 、5G、ＲＰＡ、FinTech等の最新技術への早期取り組み Tag2	, AI,, early efforts for the latest technologies such as 5G, RPA, FinTech, etc.	MT
70	Tag1 ・ Tag1 Tag2 質の高い新卒採用、中途採用の拡大、及び継続的な社員教育による高度人財の確保 Tag2		MT
71	Tag1 ・ Tag1 Tag2 優秀パートナーの開拓、既存コア・パートナーとの協業強化、及び効果的なアライアンスの実施 Tag2	, ,Develop excellent partners, strengthen collaboration with existing core partners, and implement effective alliances	MT

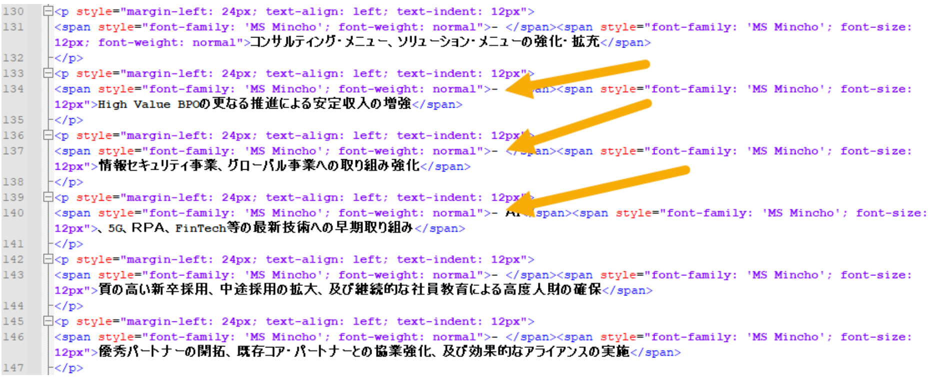
All translated segments contain mysterious commas and two segments contain open quotes. Segment 70 did not populate the first time around (the second try populated a reasonable translation with mysterious commas). None of the translations contain bullet points or dashes. Segment 69 contains a translation error caused by change in font in the source (Japanese) text (“AI” is in full-width, “5G” in half-width, “RPA” in full-width, “FinTech” in half-width)

Here’s what the file looks like after finding and replacing full-width bullet points with a half-width dash in Notepad++.



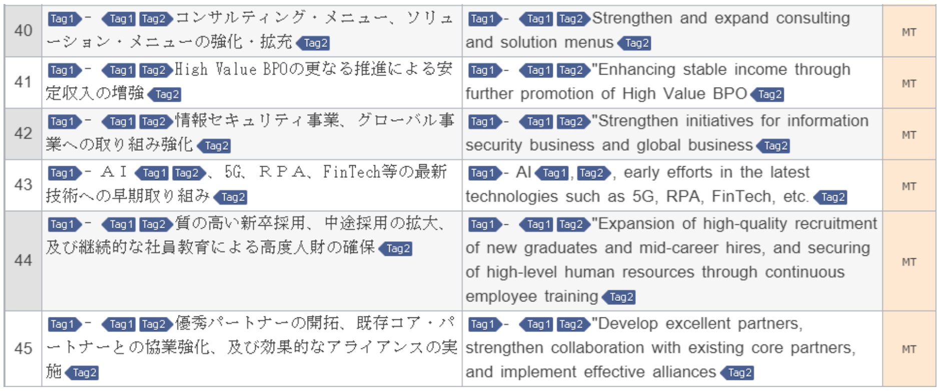
- コンサルティング・メニュー、ソリューション・メニューの強化・拡充
- High Value BPOの更なる推進による安定収入の増強
- 情報セキュリティ事業、グローバル事業への取り組み強化
- ＡＩ、5G、ＲＰＡ、FinTech等の最新技術への早期取り組み
- 質の高い新卒採用、中途採用の拡大、及び継続的な社員教育による高度人財の確保
- 優秀パートナーの開拓、既存コア・パートナーとの協業強化、及び効果的なアライアンスの実施

In Notepad++:



```
<p style="margin-left: 24px; text-align: left; text-indent: 12px">
<span style="font-family: 'MS Mincho'; font-weight: normal">- </span><span style="font-family: 'MS Mincho'; font-size: 12px; font-weight: normal">コンサルティング・メニュー、ソリューション・メニューの強化・拡充</span>
</p>
<p style="margin-left: 24px; text-align: left; text-indent: 12px">
<span style="font-family: 'MS Mincho'; font-weight: normal">- </span><span style="font-family: 'MS Mincho'; font-size: 12px">High Value BPOの更なる推進による安定収入の増強</span>
</p>
<p style="margin-left: 24px; text-align: left; text-indent: 12px">
<span style="font-family: 'MS Mincho'; font-weight: normal">- </span><span style="font-family: 'MS Mincho'; font-size: 12px">情報セキュリティ事業、グローバル事業への取り組み強化</span>
</p>
<p style="margin-left: 24px; text-align: left; text-indent: 12px">
<span style="font-family: 'MS Mincho'; font-weight: normal">- AI</span><span style="font-family: 'MS Mincho'; font-size: 12px">、5G、RPA、FinTech等の最新技術への早期取り組み</span>
</p>
<p style="margin-left: 24px; text-align: left; text-indent: 12px">
<span style="font-family: 'MS Mincho'; font-weight: normal">- </span><span style="font-family: 'MS Mincho'; font-size: 12px">質の高い新卒採用、中途採用の拡大、及び継続的な社員教育による高度人財の確保</span>
</p>
<p style="margin-left: 24px; text-align: left; text-indent: 12px">
<span style="font-family: 'MS Mincho'; font-weight: normal">- </span><span style="font-family: 'MS Mincho'; font-size: 12px">優秀パートナーの開拓、既存コア・パートナーとの協業強化、及び効果的なアライアンスの実施</span>
</p>
```

In WFP6:



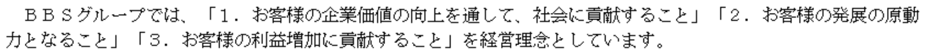
40	Tag1 - Tag1 Tag2 コンサルティング・メニュー、ソリューション・メニューの強化・拡充 Tag2	Tag1 - Tag1 Tag2 Strengthen and expand consulting and solution menus Tag2	MT
41	Tag1 - Tag1 Tag2 High Value BPOの更なる推進による安定収入の増強 Tag2	Tag1 - Tag1 Tag2 "Enhancing stable income through further promotion of High Value BPO Tag2	MT
42	Tag1 - Tag1 Tag2 情報セキュリティ事業、グローバル事業への取り組み強化 Tag2	Tag1 - Tag1 Tag2 "Strengthen initiatives for information security business and global business Tag2	MT
43	Tag1 - ＡＩ Tag1 Tag2 、5G、ＲＰＡ、FinTech等の最新技術への早期取り組み Tag2	Tag1 - AI Tag1 , Tag2 , early efforts in the latest technologies such as 5G, RPA, FinTech, etc. Tag2	MT
44	Tag1 - Tag1 Tag2 質の高い新卒採用、中途採用の拡大、及び継続的な社員教育による高度人財の確保 Tag2	Tag1 - Tag1 Tag2 "Expansion of high-quality recruitment of new graduates and mid-career hires, and securing of high-level human resources through continuous employee training Tag2	MT
45	Tag1 - Tag1 Tag2 優秀パートナーの開拓、既存コア・パートナーとの協業強化、及び効果的なアライアンスの実施 Tag2	Tag1 - Tag1 Tag2 "Develop excellent partners, strengthen collaboration with existing core partners, and implement effective alliances Tag2	MT

All translated segments contain a dash, 4 translated segments contain open quotes.
No mysterious commas. The 4 open quotes should not be there, but the error is non-critical.

2. Japanese quotes (「」) vs. English quotes (“”)

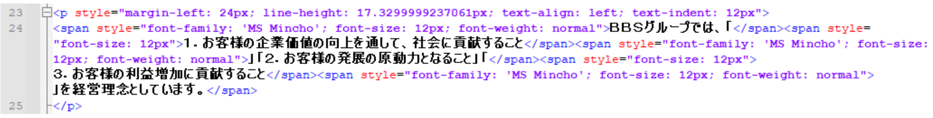
Japanese quotes are also a common cause for error for machine translation engines, probably because it is unique to the Japanese language. The tricky part here is that Japanese documents occasionally contain separate quotes in succession with no separator (i.e., a comma).

The following section found in Business Brain Showa-Ota’s FY03/20 Annual Securities Report contains Japanese quotes in succession:



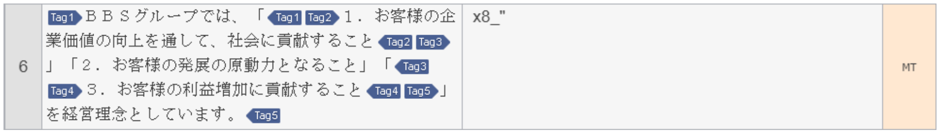
ＢＢＳグループでは、「１．お客様の企業価値の向上を通して、社会に貢献すること」「２．お客様の発展の原動力となること」「３．お客様の利益増加に貢献すること」を経営理念としています。

In Notepad++, this section appears as below:



```
<p style="margin-left: 24px; line-height: 17.3299999237061px; text-align: left; text-indent: 12px">
<span style="font-family: 'MS Mincho'; font-size: 12px; font-weight: normal">BBSグループでは、「</span><span style=
"font-size: 12px">1. お客様の企業価値の向上を通して、社会に貢献すること</span><span style="font-family: 'MS Mincho'; font-size:
12px; font-weight: normal">」「2. お客様の発展の原動力となること」「</span><span style="font-size: 12px">
3. お客様の利益増加に貢献すること</span><span style="font-family: 'MS Mincho'; font-size: 12px; font-weight: normal">
」を経営理念としています。</span>
</p>
```

Here’s a screenshot from Wordfast Pro 6 after DeepL machine translations have been populated:

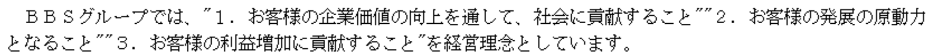


6	Tag1 ＢＢＳグループでは、「 Tag1 Tag2 １．お客様の企業価値の向上を通して、社会に貢献すること Tag2 Tag3 」「２．お客様の発展の原動力となること」「 Tag3 Tag4 ３．お客様の利益増加に貢献すること Tag4 Tag5 」を経営理念としています。 Tag5	x8_"	MT

DeepL returned a strange string of characters and no translation, making this a critical error.
This happens from time to time when dealing with too many Japanese quotes.
To fix this, here’s what I did in Notepad++:

1. Find (」　「) and replace (“”)
2. Find (「) and replace (“)
3. Find (」) and replace (“)

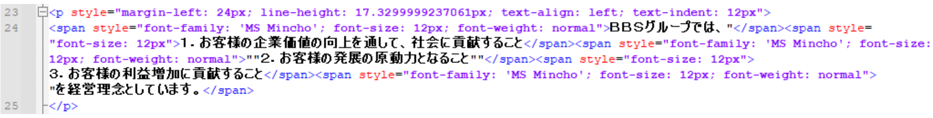
The same portion in the revised file looks like this:



ＢＢＳグループでは、″１．お客様の企業価値の向上を通して、社会に貢献すること″″２．お客様の発展の原動力となること″″３．お客様の利益増加に貢献すること″を経営理念としています。

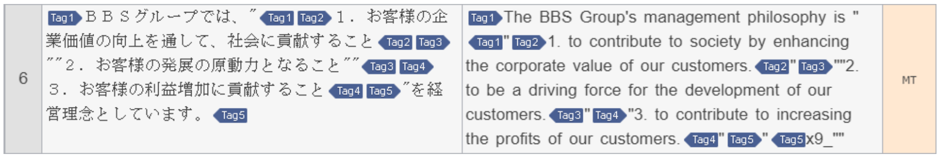
Not pretty in Japanese, but easier for computers to read.

In Notepad++, the revised section appears as below:



```
<p style="margin-left: 24px; line-height: 17.3299999237061px; text-align: left; text-indent: 12px">
<span style="font-family: 'MS Mincho'; font-size: 12px; font-weight: normal">BBSグループでは、"</span><span style=
"font-size: 12px">1. お客様の企業価値の向上を通して、社会に貢献すること</span><span style="font-family: 'MS Mincho'; font-size:
12px; font-weight: normal">""2. お客様の発展の原動力となること""</span><span style="font-size: 12px">
3. お客様の利益増加に貢献すること</span><span style="font-family: 'MS Mincho'; font-size: 12px; font-weight: normal">
"を経営理念としています。</span>
</p>
```

In WFP6:



6	Tag1 ＢＢＳグループでは、″ Tag1 Tag2 １．お客様の企業価値の向上を通して、社会に貢献すること Tag2 Tag3 ″″２．お客様の発展の原動力となること″″ Tag3 Tag4 ３．お客様の利益増加に貢献すること Tag4 Tag5 ″を経営理念としています。 Tag5	Tag1 The BBS Group's management philosophy is " Tag1 " Tag2 1. to contribute to society by enhancing the corporate value of our customers. Tag2 " Tag3 ""2. to be a driving force for the development of our customers. Tag3 " Tag4 "3. to contribute to increasing the profits of our customers. Tag4 " Tag5 " Tag5 x9_""	MT

Mysterious string of characters appeared at the end of the machine translation output, but this error is non-critical. Overall, the sentence is coherent (and accurately translated), although stylistically poor (non-critical).

The following is a non-exhaustive list of edits to consider for preprocessing. They are broadly categorized into easy edits (think: find & replace) and hard edits (think: regex or programming).

コメントの追加［1］: I've heard regex is terrible for HTML/XML editing as there are so many ways to write tags.

This package is apparently the go-to for bulk editing, if you can figure out how to use it. https://html-agility-pack.net/

Easy edits

Find	Desc.	Replace	Desc.
・	Full-width bullet point + space	-	Half-width dash + space
」　「	Japanese close quote + Japanese open quote	““	Close quote + open quote

コメントの追加［2］: Maybe replace with full-width comma + space? (、). Not sure how to handle first open quote or last close quote if we do this.

「	Japanese open quote	“	Open quote
」	Japanese close quote	“	Close quote

Hard edits

Hard returns inside table cells of XBRL files

All Japanese Annual Securities Report filings seem to have the same issue of containing hard returns within a cell, rather than wrapping the text, when the content of the cell does not fit into the designated space. This is an issue because when we feed one sentence that is broken up in pieces on different lines to the machine translation engine, the machine translation engine views each piece as a separate sentence. This in turn generates incorrect translations and messy formatting.

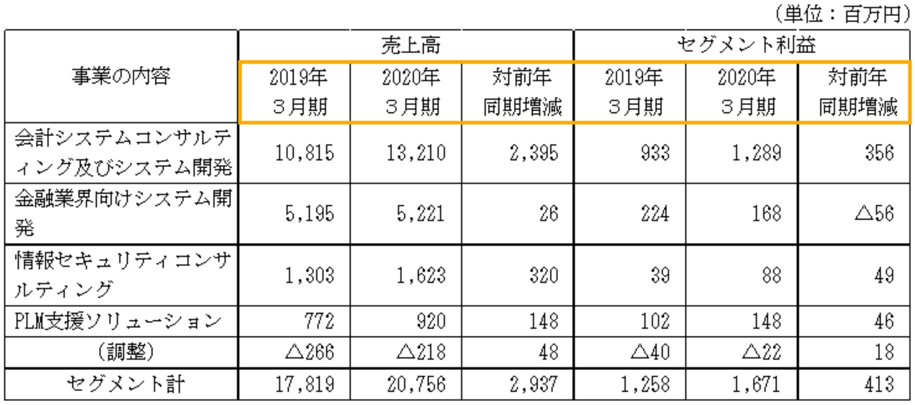
While in many cases, the fix simply involves removing all hard returns inside a cell, there are also cases where hard returns are placed to split a phrase or sentence. In the latter case, we would be concatenating sentences that should not be concatenated, which will produce translation errors.

First, we will go over a simple case.

Simple case

Here is an example from Business Brain Showa-Ota’s FY03/20 Annual Securities Report (filename: 0102010_honbun_jpcrp030000-asr-001_E04869-000_2020-03-31_01_2020-06-25_ixbrl)

The 1st table in the above file looks like this (in your browser):

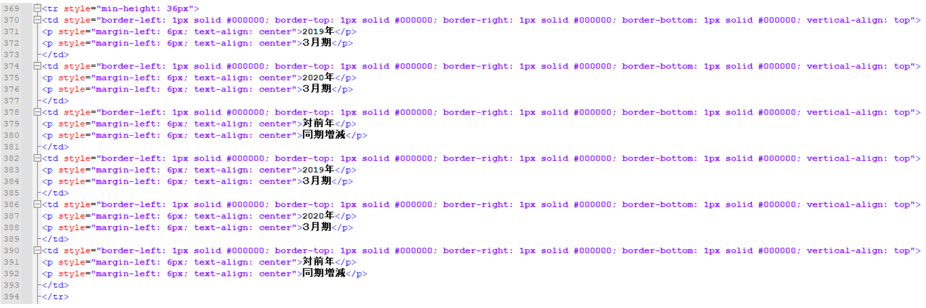


（単位：百万円）

事業の内容	売上高			セグメント利益		
	2019年 ３月期	2020年 ３月期	対前年 同期増減	2019年 ３月期	2020年 ３月期	対前年 同期増減
会計システムコンサルティング及びシステム開発	10,815	13,210	2,395	933	1,289	356
金融業界向けシステム開発	5,195	5,221	26	224	168	△56
情報セキュリティコンサルティング	1,303	1,623	320	39	88	49
PLM支援ソリューション	772	920	148	102	148	46
（調整）	△266	△218	48	△40	△22	18
セグメント計	17,819	20,756	2,937	1,258	1,671	413

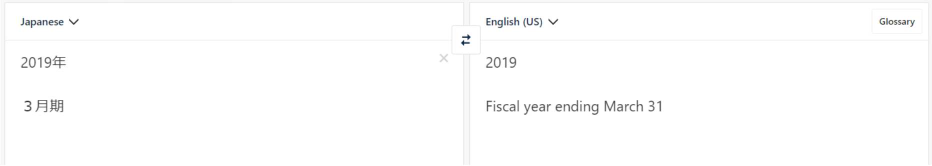
The area highlighted in orange is where we will focus on. This table shows sales and profit by business segment. The first two cells in the highlighted area (from left to right), shows the fiscal years, with the 3rd cell showing YoY change. Then the 4th and 5th cells show fiscal years, followed by YoY change in the 6th cell.

When opening the same file in Notepad++, the section for this table starts at line 348, and the area highlighted in orange starts at line 369.



```
369  <tr style="min-height: 36px">
370  <td style="border-left: 1px solid #000000; border-top: 1px solid #000000; border-right: 1px solid #000000; border-bottom: 1px solid #000000; vertical-align: top">
371   <p style="margin-left: 6px; text-align: center">2019年</p>
372   <p style="margin-left: 6px; text-align: center">3月期</p>
373  </td>
374  <td style="border-left: 1px solid #000000; border-top: 1px solid #000000; border-right: 1px solid #000000; border-bottom: 1px solid #000000; vertical-align: top">
375   <p style="margin-left: 6px; text-align: center">2020年</p>
376   <p style="margin-left: 6px; text-align: center">3月期</p>
377  </td>
378  <td style="border-left: 1px solid #000000; border-top: 1px solid #000000; border-right: 1px solid #000000; border-bottom: 1px solid #000000; vertical-align: top">
379   <p style="margin-left: 6px; text-align: center">対前年</p>
380   <p style="margin-left: 6px; text-align: center">同期増減</p>
381  </td>
382  <td style="border-left: 1px solid #000000; border-top: 1px solid #000000; border-right: 1px solid #000000; border-bottom: 1px solid #000000; vertical-align: top">
383   <p style="margin-left: 6px; text-align: center">2019年</p>
384   <p style="margin-left: 6px; text-align: center">3月期</p>
385  </td>
386  <td style="border-left: 1px solid #000000; border-top: 1px solid #000000; border-right: 1px solid #000000; border-bottom: 1px solid #000000; vertical-align: top">
387   <p style="margin-left: 6px; text-align: center">2020年</p>
388   <p style="margin-left: 6px; text-align: center">3月期</p>
389  </td>
390  <td style="border-left: 1px solid #000000; border-top: 1px solid #000000; border-right: 1px solid #000000; border-bottom: 1px solid #000000; vertical-align: top">
391   <p style="margin-left: 6px; text-align: center">対前年</p>
392   <p style="margin-left: 6px; text-align: center">同期増減</p>
393  </td>
394  </tr>
```

As you can see, there are technically two paragraphs in each cell. The machine translation engine treats each paragraph as separate entries, which means there will technically be two entries for each cell. To replicate the problem, copy the content of the first cell in your browser, then paste it into DeepL.



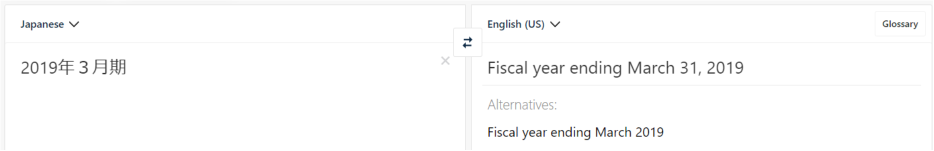
2019年

３月期

2019

Fiscal year ending March 31

To see what it’s supposed to look like, remove the hard returns and see the output:



2019年３月期

Fiscal year ending March 31, 2019

Alternatives:

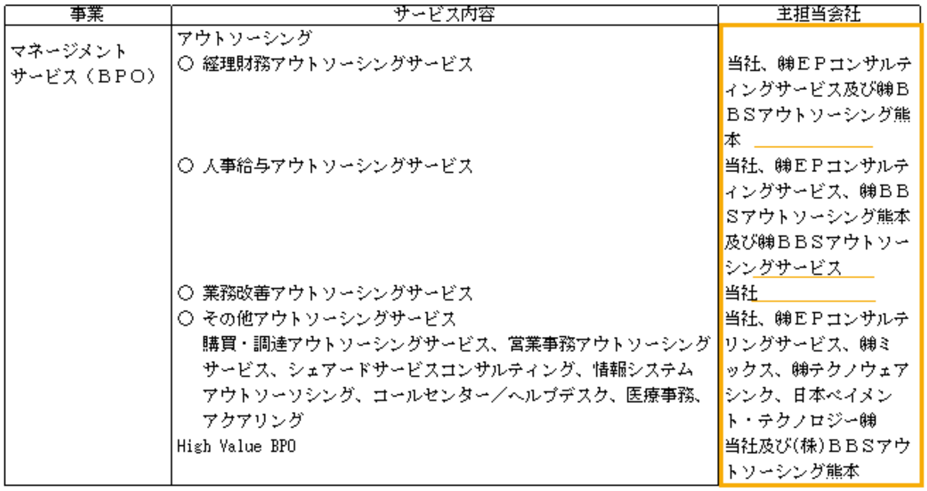
Fiscal year ending March 2019

Next, we will go over the complex case

Complex case

Here is an example from Business Brain Showa-Ota's FY03/20 Annual Securities Report (filename: 0101010_honbun_jpcrp030000-asr-001_E04869-000_2020-03-31_01_2020-06-25_ixbrl)

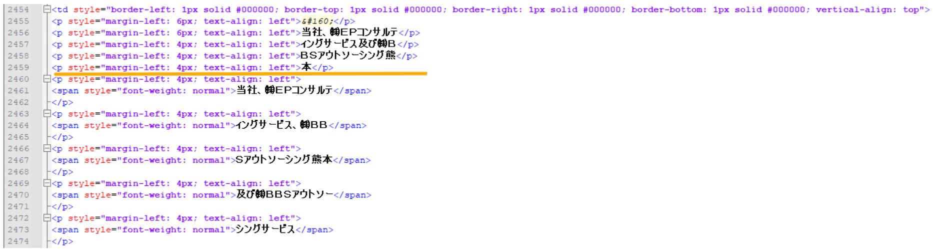
The 5th table in the above file looks like this:



事業	サービス内容	主担当会社
マネージメントサービス（ＢＰＯ）	アウトソーシング ○ 経理財務アウトソーシングサービス	当社、㈱ＥＰコンサルティングサービス及び㈱ＢＢＳアウトソーシング熊本
	○ 人事給与アウトソーシングサービス	当社、㈱ＥＰコンサルティングサービス、㈱ＢＢＳアウトソーシング熊本及び㈱ＢＢＳアウトソーシングサービス
	○ 業務改善アウトソーシングサービス	当社
	○ その他アウトソーシングサービス 購買・調達アウトソーシングサービス、営業事務アウトソーシングサービス、シェアードサービスコンサルティング、情報システムアウトソーシング、コールセンター／ヘルプデスク、医療事務、アクアリング	当社、㈱ＥＰコンサルリングサービス、㈱ミックス、㈱テクノウェアシンク、日本ペイメント・テクノロジー㈱
	High Value BPO	当社及び(株)ＢＢＳアウトソーシング熊本

The area highlighted in orange is where we will focus on. The orange lines within the cell are places where hard returns are actually necessary.

When opening the same file in Notepad++, the section for this table starts at line 2,389, and the area highlighted in orange starts at line 2,454.



```
<td style="border-left: 1px solid #000000; border-top: 1px solid #000000; border-right: 1px solid #000000; border-bottom: 1px solid #000000; vertical-align: top">
 <p style="margin-left: 6px; text-align: left">&#160;</p>
 <p style="margin-left: 6px; text-align: left">当社、㈱ＥＰコンサルテ</p>
 <p style="margin-left: 4px; text-align: left">ィングサービス及び㈱Ｂ</p>
 <p style="margin-left: 4px; text-align: left">ＢＳアウトソーシング熊</p>
 <p style="margin-left: 4px; text-align: left">本</p>
<p style="margin-left: 4px; text-align: left">
 <span style="font-weight: normal">当社、㈱ＥＰコンサルテ</span>
</p>
<p style="margin-left: 4px; text-align: left">
 <span style="font-weight: normal">ィングサービス、㈱ＢＢ</span>
</p>
<p style="margin-left: 4px; text-align: left">
 <span style="font-weight: normal">Ｓアウトソーシング熊本</span>
</p>
<p style="margin-left: 4px; text-align: left">
 <span style="font-weight: normal">及び㈱ＢＢＳアウトソー</span>
</p>
<p style="margin-left: 4px; text-align: left">
 <span style="font-weight: normal">シングサービス</span>
</p>
```

The orange line is where the hard return is actually necessary. I have no idea how to approach this issue and there is seemingly no way to programmatically distinguish where a hard return is necessary/unnecessary.

One roundabout way of approaching this might be to capture the width of the cell (from <thead> element), estimate how many characters would fit in a single line based on font style (full-width vs. half-width) and font size, then for each hard return, check whether the previous line is filled edge to edge. Presumably, if the previous line is filled edge to edge, the line following will be a continuation of the previous line. If the previous line is not filled edge to edge, then it’s highly likely that the following line is a new phrase and not a continuation of the previous line.