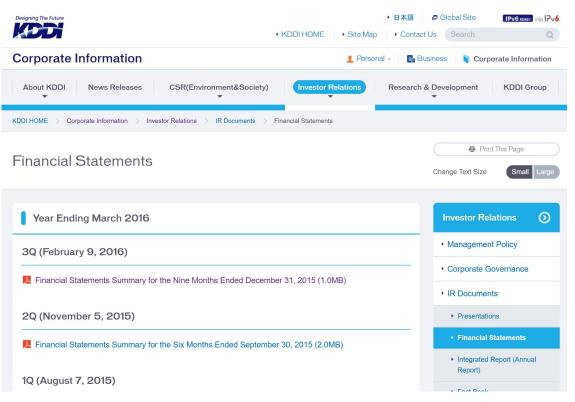
BESST MT Pilot Project

Ben, Emily, Sam, Shuwen Team

Background

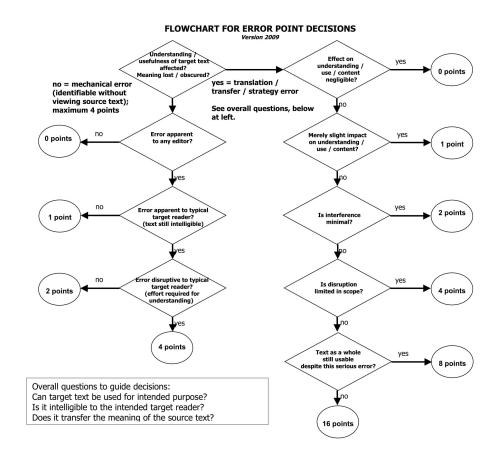
Financial Statements, Japanese to English

- KDDI Financial Statements
 - o 2003 **-** 2015
- Softbank
 - o **1999 2015**
- Docomo
 - 0 1998 2015



Initial Goals

- Quality
 - Almost good enough for ATA
- Efficiency
 - \circ ~~ 70% faster than human translation
- Cost
 - \circ ~~ 70% cheaper than human translation



Machine Training Chart

1	0	Training: 2010 - 2015 Softbank financial reports (except for 2012q3 and 2015q4 - problem with files) pdfs: 10,670 segments Tuning: AU's Financial Statements for Years Ending March 2014 and 2015 (8 docs) pdfs: 2,731 segments Testing: AU's Financial Statements for Year Ending March 2016 (3 docs) pdfs: 2,013 segments	14.36	Emily
2	1	Added Softbank annual reports for 2015 and 2014	14.66	Sam
3	2	Added 8 KDDI financial statements	14.63	Ben
4	2	Training: Added Softbank financial reports for FY 2009 and FY2008 (unselected those with high attention marks) Tuning: Softbank and KDDI (those with higher similarity in parellel sentece)	10.75	Shuwen
5	2	Training: No change Tuning: Converted pdfs of AU 2015 year end statement to RTF and cleaned up files (removed images, headers, footers, and sections of English translation not found in original). Also removed all numbers and most symbols. Ran files through aligner (used MemoQ) and then cleaned alignment for ~2 hours.	14.05	Emily
6	2	Added Docomo Annual Report HTML files	16.85	Sam
7	6	Realized that our testing data was pretty awful so I tried to run some OCR on the Japanes PDFs in order to have all of the text picked up and actually represented in words/sentences. Before, it looked like there was a space in between every single character in the Japanese. I only did it on one of the documents, but we'll see if it helps at all.	18.18	Ben
8	7	Training: SB doc files, html, tmx Tuning: KDDI doc files	15.51	Shuwen
9	7	Training : added monolingual docs (2015 10k reports from Verizon, Sprint, T-Mobile, and AT&T)	19.03	Emily
10	9	More of what I did before with trying to clean up our testing data, because it wasn't very clean, this wont be perfect but it should be a lot better than what it was before. Its the same "data" just hopefully easier to segment and hopefully provide better results	22.03	Ben



Initial BLEU score : 14.36

Best BLEU score: 22.03 (53% increase)

Things that worked: OCR, HTML files, cleaning out trash, monolingual docs

Things that didn't: PDFs, and trying to align them

Goal Evaluation

- Quality
 - $\circ \quad \text{Not quite within our quality metrics} \\$
- Efficiency
 - \circ Based on our initial pilot project and various assumptions
 - Approximately 85% (Goal was 70%)
- Price
 - Based on what we would be charging the customer/vendors, doesn't include training costs
 - Approximately 67% (Goal was 70%)
 - Depends on word count required to "break-even" based on training costs



📓 C:\Users\Benjamin\Desktop\Sentences_AU+Financial+Statements_ja_en_20160426 (1)\Align 🗕 🗖	
<u>File Edit Search View Encoding Language Settings Macro Run Plugins Window 2</u>	x
📙 kddi_2016_3qc_vN19Wi_ja.align 🛛 🗒 updated_kddi_2016_3qc_vN19Wi_ja.align 🗵	
18 また、共同住宅や難視聴対策としても利用される。 19 さらに、CATV 用のケーブルを使って、テレビ放送ばかりでなく、インターネットや電話の利用も可能となっている。 20 FTTHLは、Fiber To The Homeの略。	A Benjamin\Desktop\Sentences_AU+Financial+Statements_ja_en_20160426\Aligned\ -
21 通信事業者の設備からお客様宅までを光ファイバーケーブルでつなぐアクセス方式。 22 本来は「Home」のとおり個人のお客様を対象としたものであるが、光ファイバーによるアクセス全般を指 すこともある。	a 🕞 🕞 🔐 🐇 🐂 🐂 🥥 ⊄ 🇰 ½ 🔍 👒 🖳 🚭 🔤 11 🏣 🐼 📡 / 🔤 💌 💌 ា 🕨 🐚 qc_vN19Wi_js_align 🛛 🖶 updated_kddi_2016_3qc_vN19Wi_js_align 💟
23 ICT ICTCは、Information and Communication Technologies:ICTs(情報通信技術)の略。 24 以前は、ITCいう言葉が主に使われたがインターネット時代の昨今では、コ	「Home」のとおり個人のお客様を対象 としたものであるが、光ア・イバーによるアクセス全般を指すこともある。 Tとは、Information and Communication Technologies:ICTs (情報通信技術)の 略。
ンピュータシステムのみならず、さまざまなシステムが通信ネットワークで接続されることによって、多くの付加価値を 産み出すことから、「ICT」の語が用いられることが多くなった。	、ITという言葉が主に使われたが、インターネット時代の昨今では、コ -タシステムのみならず、さまざまなシステムが通信ネットワークで接続さ
25 LTE LTEとは、Long Term Evolutionの略称で、無線通信技術の一つ。 26 MVNO MVNOとは、Mobile Virtual Network Operator(仮想移動体サビス事業者)の略。	こによって、多くの付加価値を産み出すことから、「ICT」の語が用いられることが多くなった。 TEEとは、Long Term Evolutionの略称で、無線通信技術の一つ。
27 無線通信 インフラを他携帯電話事業者等から借り受けてサービスを提供している事業者のこと。 28 VOLTE WIMAX 2+ VOLTEとは、Voice over LTEの略称で、高速通信規格「LTE」を利用した音 声通話技術のこと。	MVNOとは、Mobile Virtual Network Operator (仮想移動体サ1ービス事業者)の略。 M信インフラを他携帯電話事業者等から借り受けてサービスを提供している事 業者のこと。 EVOLTBとは、Voice over LTEの略称で、高速通信規格[LTE1を利用した音声通話技 術のこと。
戸通語 校 M 00-C。 29 LTEO 利用 (L-U) クリアで聞き取りやすい 音声通話が可能となる。 30 また、音声通話中 (L-Web 閲覧やメールの送受信を行うなど、音声通話とデータ通信が同時	利用により、クリアで聞き取りやすい音声通話が可能となる。 音声通話中に叱ゃ閲覧やメールの送受信を行うなと、音声通話とデータ通信が同時に利用できる。
30 に利用できる。 31 WiMAX 2+とは、UQIミュニケーションズ株式会社が提供するBroadband Wireless Accessサービスのひとつ。	x 2+WiMAX 2+とは、UQコミニケ1ーションズ株式会社が提供するBroadband Wireless Accessサービスのひとつ。 DモバイルwiMAXに比べ、周波数の利用効率が高く
32 従来のもいいWiMaxに比べ、周波数の利用効率が高く、高速通信が可能で2.6GHz帯 の20MHz幅を使い、受信最大速度110Mbps、送信最大速度10Mbpsを実現する。	
33 誌、[TD-LTE]方式との互換性を有する。 34 なわ、2015 年春がは、2.66Hz 帯の20MHz 幅を2つ束ねるcaを使い、受信時最大	2015年春 からは、2.6日z帯の20HHz幅を2つ束ねるCAを使い、受信時最大220Mbpsのサービス を提供している。 通信事業者は新たな収益の確保に向けて通信以外の
34 /a0,2013 年 日 かるに 2.00m2 冊 020m2 冊 227 来 aloca 度 いく 文 日 時 最 八 220Mbpsのサービスを 提 供 している。 25 日 木 其 進 かみてEPS 其 進 へ 変 軍 するにあたっての 調 整 の 詳 細 は P 45[3	スへ事業領域を拡大しつつあり、各社の事業戦略は大きな転換期を迎えております。
Normal text file length : 94586 lines : 405 Ln : 24 Col : 29 Sel : 0 0 Dos/Windows UCS-2 LE BOM IN	
	としないです来な地域のようにしたいのし、紙をプログラインではないができた。 各目指していために、「3M戦略」と「グローバル戦略」を推進しております。 当第3四半期には「au WALLET Market」の全国展開や電力 小売事業への参入決定をしました。
76	日本3回半期には「du WALLET Market」の全国展開「電力」「元事素への多人大足どしなし。 誌、昨年12月3日には「au WALLET Market」の会員数が100万 人を突破しました。 第3四半期車注思計期間(単位・百万円)
Normal	

Recommendations

- Ask the client directly for source files
 - Ideally some form of inDesign or Microsoft Word document
 - Not PDF
- Adding more documents/segments to our training/tuning set
 - Find more related companies and their financial reports
- Alignment, lots of alignment
 - We did do some alignment but a large majority of our documents had alignment issues, so we were unable to do very many in the time available to us.
 - Should drastically improve results, but will take a lot of time and manpower
- Adding more monolingual texts and dictionaries
 - Specifically a dictionary of financial related terms that would likely appear in an annual report or other related financial documents.

Timeline/Costs

30-Mar	30-Apr	30-May	30-Jun	30-Jul
• Kickoff				
\checkmark				
Ma	chine trair	ning		
		\checkmark		
			Post editi	ng
				\downarrow
				Project
				completed

Task	Estimated Hours	Hourly Rate	Cost
Document alignment	40	\$40.00	\$1,600
MT training	12	\$30.00	\$360
Post-editing	8	\$30.00	\$240
QA	4	\$30.00	\$120
Total estimated project cost			

Lessons Learned

- Need to have easily accessible data (and lots of it)
- Training: Quantity > Quality
 - Use tools (eg, Olifant) not manpower to clean up
- Testing: Quality > Quantity
- Expert help
- One change at a time

(百万円未満四捨五入)	Ref: (Amounts are rounded off to nearest million yen) MT: (Rounded to the nearest million yen)
(1)連結経営成績(累計)(%表示は、対前年同四半期増減率)	Ref: COnsOlidated Operating Results(Percentage represents comparison change to the corresponding previous quarterly period) MT: (1) consolidated operating results (cumulative) (percentage represents comparison to previous fiscal year recorded in the same period decrease)
売上高営業利益税引前利益四半期利益親会社の戸:T期If#屑する四四半期鰭利益合	Ref: OperatingrevenueOperatingincomeProfit for the period before income taxProfit for the periodProfit for the period attributable to owners of the parentTotalcomprehensive income for the period MT: Door: period If of revenues operating income tax income before income quarter # waste to 44 quarter fin income total
百方円%百万円%百万円%百万円%百万円%	Ref: Nine-month period%%%%%% MT: Million yen % million: million yen million yen % million yen million yen %.
28年3月期第3四半期3,299,0313.8672,44211.0662,3709.0456,03216.9408,48613.4449,2439.6	Ref: ended December 31, 2015Nine-month period3,299,031 3.8672,44211.0662,370 9.0456,032 16.9408,486 13.4449,243 9.6 MT: 3/28 period no. 3 quarter 3299, 0313.8672 44211.0662 3709.0456, 03216.9408, 48613.4449, 2439.6
27年3月期第3四半期3,178,545—605,989—607,816—390,162—360,340—409,881—	Ref: ended December 31, 20143,178,545 -605,989-607,816 -390,162 -360,340 -409,881 - MT: 3/27 period no. 3 quarter 3178, 545 — 605989 — 607816 — 390, 162-360, 340 — 409881 —
基本的1株当たy四半期利益希薄化後1株当たy四半期利益	Ref: Basic earnings per shareDiluted earnings per share MT: Basic share this was y quarter income diluted shares into the y income
28年3月期第3四半期163.04-	Ref: Nine-month period ended December 31,2015163.04- MT: 3/28 period no. 3 quarter 163.04 as
27年3月期第3四半期143.85—	Ref: Nine-month period ended December 31,2014143.85- MT: 3/27 period no. 3 quarter 143.85 —
(2)連結財政状態	Ref: (2) Consolidated Financial Positions MT: (2) consolidated financial position

Questions?