Building Salesforce Neural Machine Translation System

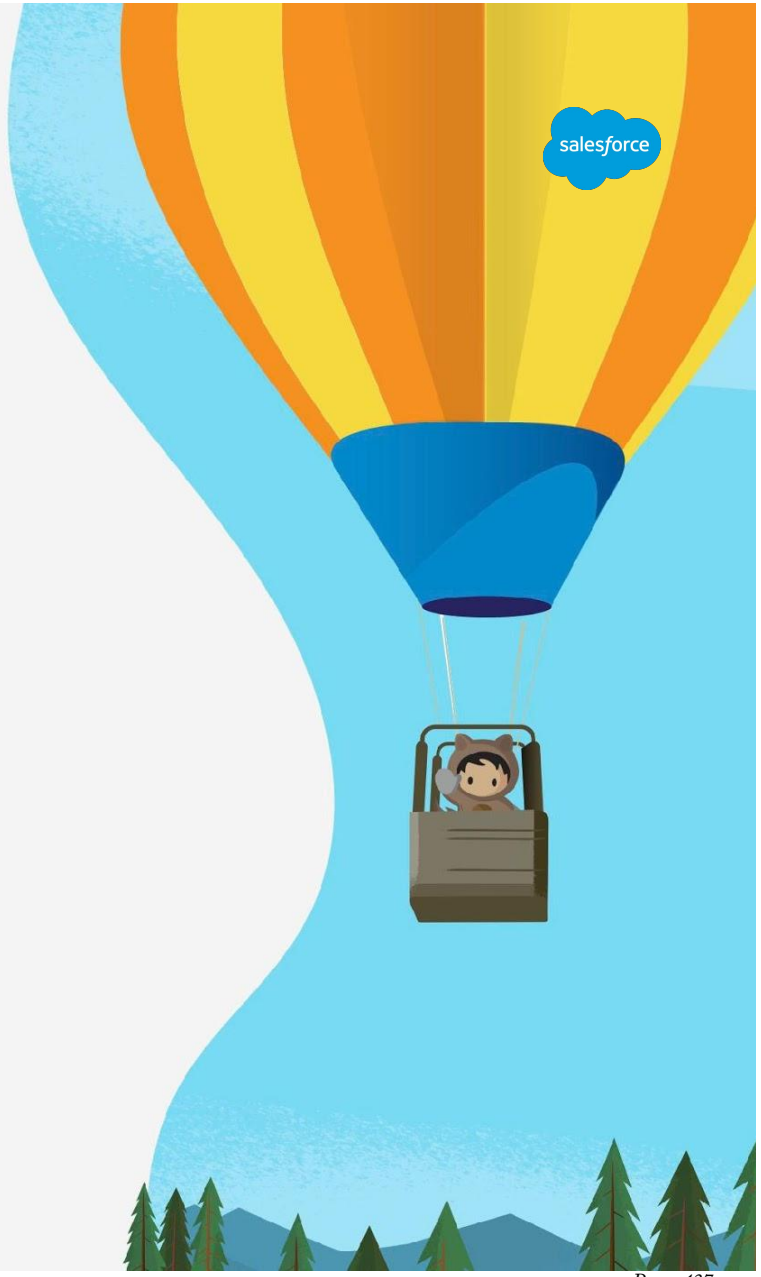
Kazuma Hashimoto, Lead Research Scientist
@ Salesforce Research

Raffaella Buschiazzo, Director, Localization
@ Salesforce R&D Localization

AMTA 2020 Commercial Track

Agenda

- Why invest in machine translation
- Salesforce online help
- What was done: Phase I
  - Technical overview
  - Example flows
- What was done: Phase II
- Roadmap



Why Invest in Machine Translation

A three-year collaboration between R&D Localization and Salesforce Research teams

Interesting research project

- Challenges: difficult MT languages (i.e. Finnish, Japanese), XML tagging.

Improve international customer experience by

- Reducing translation time by enhancing translator's productivity for our online help
- Increasing content accuracy/freshness by publishing updates more frequently
- Re-investing savings into high-value efforts
  - Products and product-related properties
  - Underserved localization content/efforts

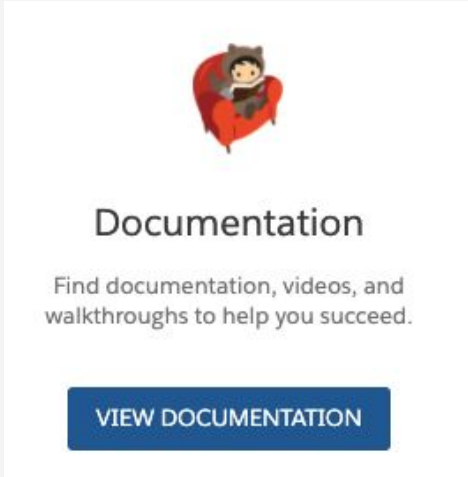
Benefits

- Increase case deflection through up-to-date content for existing languages
- Increase breadth and depth of localization coverage with more flexibility by market

Salesforce Online Help

Primary target for our MT system

- Translated in 16 languages.
- Translations are updated per major release (3 x year).
- New feature/product terminology.
- Structured in DITA XML (200+ tags).



English
Français
Deutsch
Italiano
日本語
Español (México)
Español
中文（简体）
中文（繁體）
한국어
Русский
Português (Brasil)
Suomi
Dansk
Svenska
Nederlands
ภาษาไทย
Norsk

What Was Done: Phase I

Linguistic testing

Built an NMT system on Salesforce domain

- Language-agnostic architecture with models for each language
- Processes whole XML files from English into 16 languages

Completed human evaluations of MTed output

- Japanese, Finnish, German, French Help subsets (500 strings)

Published paper A High-Quality Multilingual Dataset for Structured Documentation Translation (WMT 2019)

Technical Overview

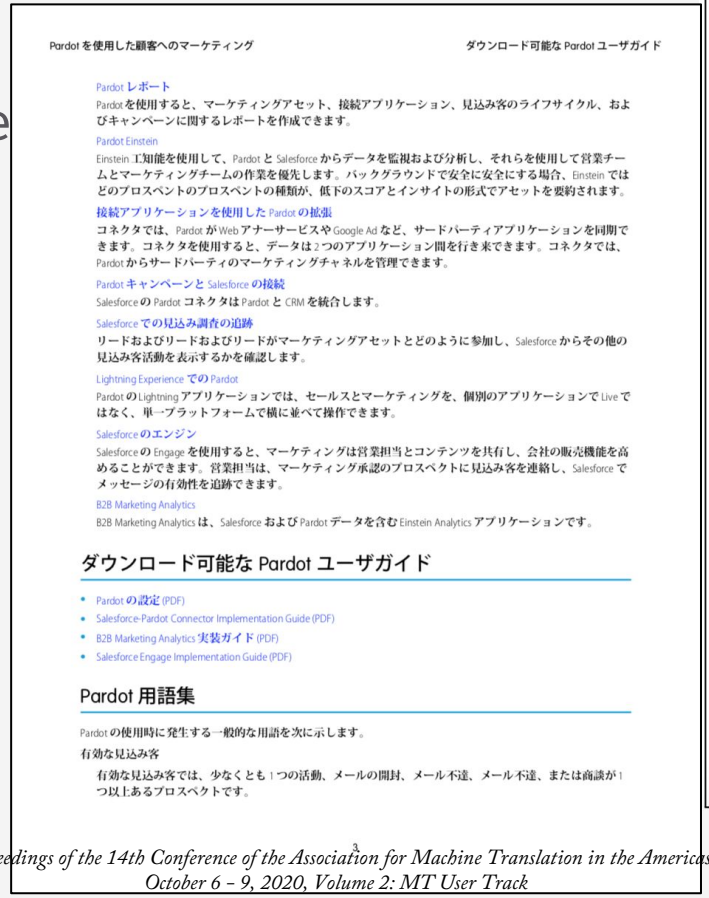
Data and application

Dataset in our paper

- https://github.com/salesforce/localization-xml-mt

Translation of rich-formatted text

- How to preserve the structure



Pardot を使用した顧客へのマーケティング　　ダウンロード可能な Pardot ユーザガイド

Pardot レポート
Pardot を使用すると、マーケティングアセット、接続アプリケーション、見込み客のライフサイクル、およびキャンペーンに関するレポートを作成できます。

Pardot Einstein
Einstein 工知能を使用して、Pardot と Salesforce からデータを監視および分析し、それらを使用して営業チームとマーケティングチームの作業を優先します。バックグラウンドで安全に安全にする場合、Einstein ではどのプロスペントのプロスペントの種類が、低下のスコアとインサイトの形式でアセットを要約されます。

接続アプリケーションを使用した Pardot の拡張
コネクタでは、Pardot が Web アナーサービスや Google Ad など、サードパーティアプリケーションを同期できます。コネクタを使用すると、データは 2 つのアプリケーション間を行き来できます。コネクタでは、Pardot からサードパーティのマーケティングチャネルを管理できます。

Pardot キャンペーンと Salesforce の接続
Salesforce の Pardot コネクタは Pardot と CRM を統合します。

Salesforce での見込み調査の追跡
リードおよびリードおよびリードがマーケティングアセットとどのように参加し、Salesforce からその他の見込み客活動を表示するかを確認します。

Lightning Experience での Pardot
Pardot の Lightning アプリケーションでは、セールスとマーケティングを、個別のアプリケーションで Live ではなく、単一プラットフォームで横に並べて操作できます。

Salesforce のエンジン
Salesforce の Engage を使用すると、マーケティングは営業担当とコンテンツを共有し、会社の販売機能を高めることができます。営業担当は、マーケティング承認のプロスペクトに見込み客を連絡し、Salesforce でメッセージの有効性を追跡できます。

B2B Marketing Analytics
B2B Marketing Analytics は、Salesforce および Pardot データを含む Einstein Analytics アプリケーションです。

ダウンロード可能な Pardot ユーザガイド

- Pardot の設定 (PDF)
- Salesforce-Pardot Connector Implementation Guide (PDF)
- B2B Marketing Analytics 実装ガイド (PDF)
- Salesforce Engage Implementation Guide (PDF)

Pardot 用語集

Pardot の使用時に発生する一般的な用語を次に示します。

有効な見込み客
有効な見込み客では、少なくとも 1 つの活動、メールの開封、メール不達、メール不達、または商談が 1 つ以上あるプロスペクトです。

- Example (a)

English:
You can use this report on your Community Management Home dashboard or in <ph>Community Workspaces</ph> under <menucascade><uicontrol>Dashboards</uicontrol><uicontrol>Home</uicontrol></menucascade>.

Japanese:
このレポートは、[コミュニティ管理] のホームのダッシュボード、または <ph>コミュニティワークスペース</ph>の <menucascade><uicontrol>[ダッシュボード]</uicontrol> <uicontrol>[ホーム]</uicontrol></menucascade> で使用できます。

- Example (b)

English:
Results with bothbeach and house in the searchable fields of the record.

Japanese:
レコードの検索可能な項目に beach と house の 両方が含まれている結果。

- Example (c)

English:
You can only predefine **this field** to an email address. You can predefine **it** using either T (used to define email addresses) or To Recipients (used to define contact, lead, and user IDs).

Japanese:
この項目はメールアドレスに対してのみ事前に定義できます。
この項目は [宛先] (メールアドレスを定義するために使用) または [宛先受信者] (取引先責任者、リード、ユーザ ID を定義するために使用) のいずれかを使用して事前に定義できます。

Technical Overview

Model

Transformer encoder-decoder (Vaswani et al., 2017)

- Input: XML-tagged text in English
- Output: XML-tagged text in another language
  - **XML-tag-aware tokenizer** is used (based on sentencepiece)
  - e.g.) <uicontrol>New Suite</uicontrol>: Create a suite of test classes that...
    → _ <uicontrol> New _Suite </uicontrol> : _Create _a _suit e _of _test _classes _that...
- + copy mechanisms
  - Copy from source is used to **align XML tags**

- Source to be translated (English)
<xref>View a single feed update</xref> by clicking the timestamp below the update, *for example*, <uicontrol>Yesterday at 12:57 AM</uicontrol>.

- Retrieved source (English)
In a feed, click the timestamp that appears below the post, *for example*, <uicontrol>Yesterday at 12:57 AM</uicontrol>.

- Retrieved reference (Japanese)
フィード内で、*たとえば*、<uicontrol>[昨日の 12:57 AM]</uicontrol> のように、投稿の下に表示されるタイムスタンプをクリックします。

- Output of the X*rs* model (Japanese)
<uicontrol> [昨日の 12:57 AM] </uicontrol> のように、更新の下にタイムスタンプをクリックして、<xref> 1 つのフィード更新を表示</xref>します。

Technical Overview

System

Training

- Construct our training data from
  - the ***N*-th** release
    - a later version than our published dataset
  - release notes of the new, **(N+1)-th**, release
    - to incorporate translation of new features/context in the new release
    - available for our company's top-tier languages
  - [optional and if applicable] whatever internal parallel data

Translation

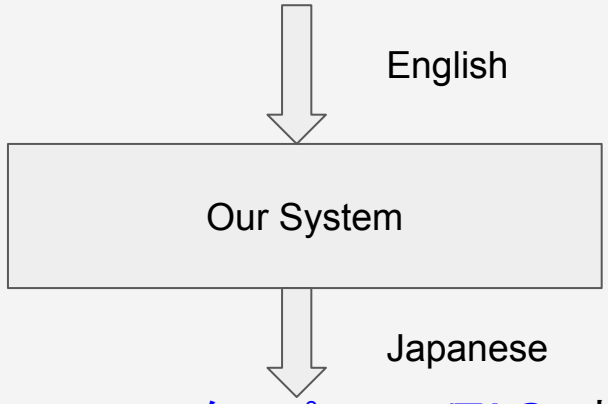
- Target English strings that have **little overlap** with our translation memory
- Remove metadata from XML tags
- Run our model for each language
- Align the metadata with the translated strings by using our model's copy mechanism

Human verification and post-editing before publishing the translated online help

Example Flow (1)

Overview

Update basic community settings like your community URL, community name, members, login options, and general preferences in the <TAG id="1">Administration</TAG> section of <TAG id="2">Experience Workspaces</TAG> or <TAG id="3">Community Management</TAG>.

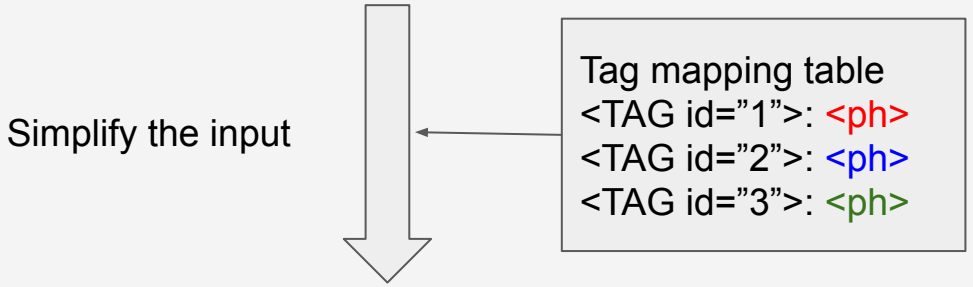


<TAG id="2">エクスペリエンスワークスペース</TAG>または <TAG id="3">[コミュニティ管理]</TAG> の <TAG id="1">[管理]</TAG> セクションで、コミュニティ URL、コミュニティ名、メンバー、ログインオプション、一般的な設定など、コミュニティの基本設定を更新します。

Example Flow (2)

Input Preprocessing

Update basic community settings like your community URL, community name, members, login options, and general preferences in the <TAG id="1">Administration</TAG> section of <TAG id="2">Experience Workspaces</TAG> or <TAG id="3">Community Management</TAG>.



Update basic community settings like your community URL, community name, members, login options, and general preferences in the <ph>Administration</ph> section of <ph>Experience Workspaces</ph> or <ph>Community Management</ph>.

Example Flow (3)

Translation by our model

Update basic community settings like your community URL, community name, members, login options, and general preferences in the <ph>Administration</ph> section of <ph>Experience Workspaces</ph> or <ph>Community Management</ph>.

<ph>エクスペリエンスワークスペース</ph>または <ph>[コミュニティ管理]</ph> の <ph>[管理]</ph> セクションで、コミュニティ URL、コミュニティ名、メンバー、ログインオプション、一般的な設定など、コミュニティの基本設定を更新します。

Example Flow (4)

Tag Alignment

Update basic community settings like your community URL, community name, members, login options, and general preferences in the <ph>Administration</ph> section of <ph>Experience Workspaces</ph> or <ph>Community Management</ph>.

Maximize the product of the copy weights based on one-to-one mapping assumption

English \ Japanese	<ph>_ja	<ph>_ja	<ph>_ja
<ph>_en	0.01	0.05	**0.91**
<ph>_en	**0.92**	0.02	0.01
<ph>_en	0.01	**0.95**	0.01

<ph>エクスペリエンスワークスペース</ph>または <ph>[コミュニティ管理]</ph> の <ph>[管理]</ph> セクションで、コミュニティ URL、コミュニティ名、メンバー、ログインオプション、一般的な設定など、コミュニティの基本設定を更新します。

Example Flow (5)

Output Postprocessing

<ph>エクスペリエンスワークスペース</ph>または <ph>[コミュニティ管理]</ph> の <ph>[管理]</ph> セクションで、コミュニティ URL、コミュニティ名、メンバー、ログインオプション、一般的な設定など、コミュニティの基本設定を更新します。

Tag mapping table
<TAG id="1">: <ph>
<TAG id="2">: <ph>
<TAG id="3">: <ph>

<TAG id="2">エクスペリエンスワークスペース</TAG>または <TAG id="3">[コミュニティ管理]</TAG> の <TAG id="1">[管理]</TAG> セクションで、コミュニティ URL、コミュニティ名、メンバー、ログインオプション、一般的な設定など、コミュニティの基本設定を更新します。

What Was Done: Phase II

Completed 2 pilots

- MTPEd two major releases of help content in Japanese, French, German, Brazilian Portuguese, Mexican Spanish, Swedish, Danish, Norwegian.

Evaluated 500 strings: our system against uncustomized commercially available NMT system

Observations:

- Salesforce NMT is better at outputting sentences with Salesforce writing style.
- Other system is good at outputting generally well-written sentences.
- Most challenging part is translating new features/terminology.
- Including Salesforce Release Notes in training data increased score #1.

Roadmap

- Leveraging publicly available models
  - So far, we used our own data only
  - Fine-tune/customize general models/engines
    - Publicly available pretrained models: mBART, XLM-R, etc.
- Human-in-the-loop training
  - At every release, we can get post-edited strings
  - Can we use the feedback to train another model to refine MT output?
    - Or can we train a model to spot potentially wrong segments to help human post-editing?
- Continual learning
- Extend MT to more online languages and more use cases


thank you
BLAZE YOUR TRAIL
salesforce