INVISIBLE MT

Patricia O'Neill-Brown, Ph.D
AMTA 2016

Overview

1. The Goal: Better Quality Machine Translation

2. 'Invisible MT'….a way to advance the field

3. Technical Approach

4. Questions/Discussion

INVISIBLE MT

MT processes today…

1. See paragraphs or documents you want to translate
2. Cut text
3. Open MT application
4. Paste into app
5. Wait for system to translate
6. See output & try to read/make sense of it
7. Decide what to do next
8. Maybe nothing else because your information need was met OR
9. Nothing else if your information need wasn't met OR
10. Post-edit - keep some output, discard some, revise still yet others

Current MT Paradigm

- Assumptions about **inputs to the system:**
  - Sentences (grammatical)

- The **output**:
  - Each system produces only one translation;
  - There is only ever one right answer

- Task **accomplishment** mode:
  - Complete Automation
  - Send document, get result

However…

What if… we changed the process & the paradigm?

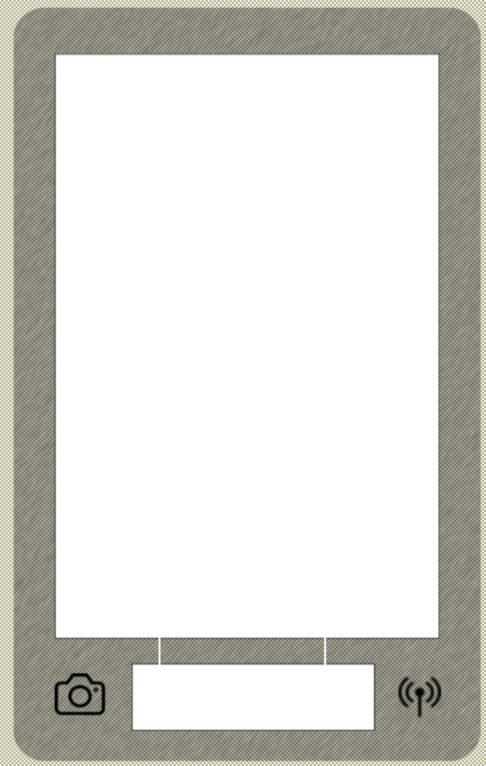
- *What if* we designed for the input being:
  - A few short sentences
  - Fragments
  - Single words
  - Images, emoticons

- What if for the **output**, we said:
  - The system's goal is to convey meaning;
  - Translation is one technique & you can use others;
  - There can be multiple 'right' ways of saying things.

- How about for the task **accomplishment** mode, it were designed as:
  - Human-in-the-MT-loop

Our Concept

Invisible MT



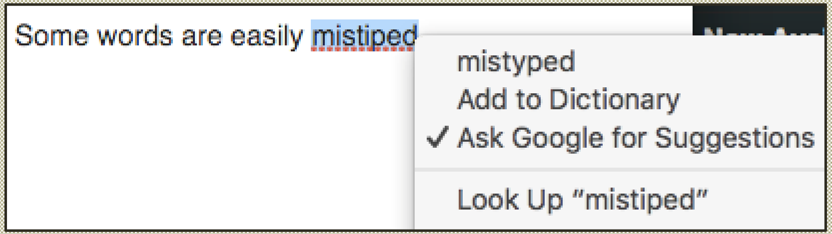
Invisible MT…

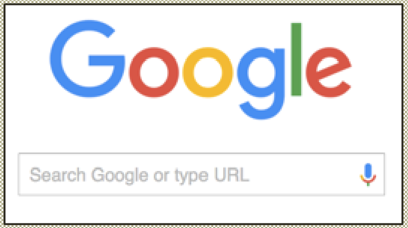
The design is driven by the *form factor* of the hardware used &
the way people are used to interacting with it already

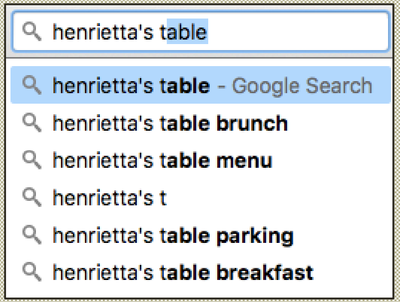
Invisible MT…

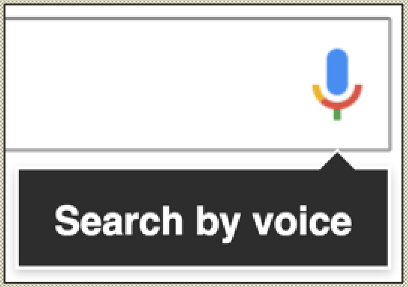
The design is also driven by the *characteristics* of the data for the different use cases

‘SEARCH’ TECHNOLOGY IS *INVISIBLE* TO THE USER









Autocomplete

Predictive Typing

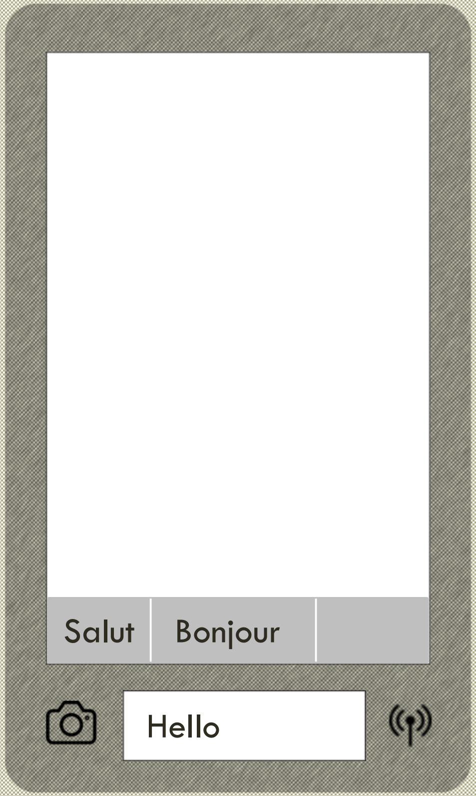
Spelling Correction

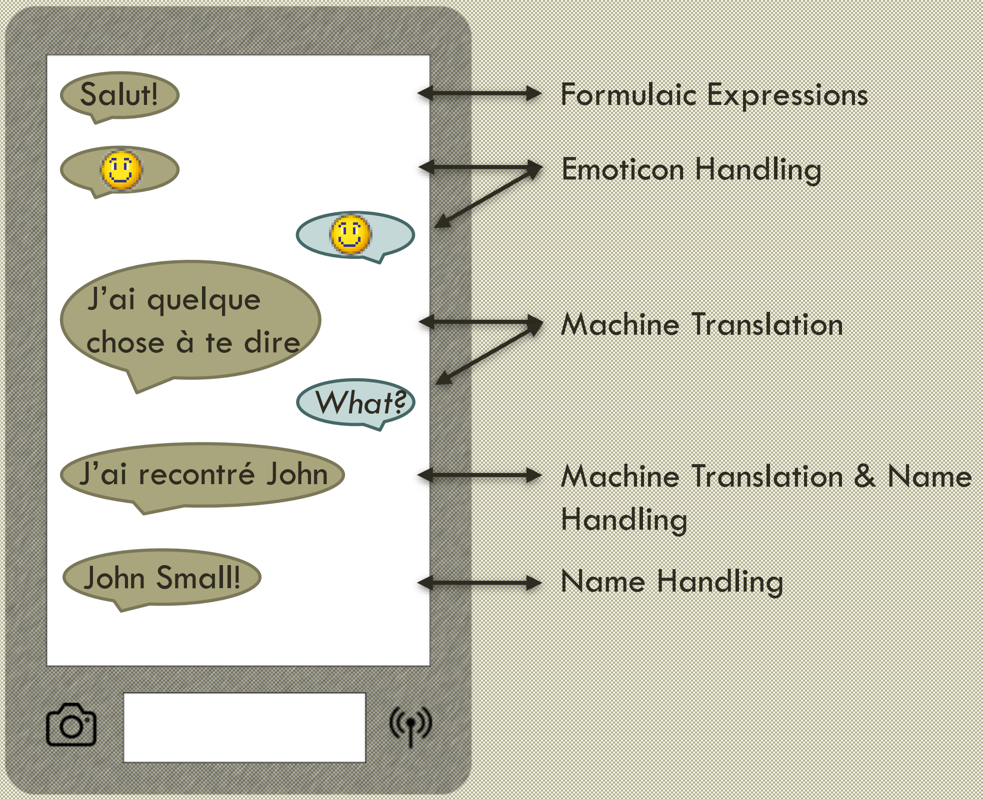
Voice Authentication

Information Retrieval

Interactive Voice Response

Automatic Speech Recognition


Salut
Bonjour
Hello


Salut!
J'ai quelque chose à te dire
What?
J'ai recontré John
John Small!
Formulaic Expressions
Emoticon Handling
Machine Translation
Machine Translation & Name Handling
Name Handling

Non-Invasive

Invisible MT

Natural

'autosuggest'

Easy

No having to go to another app

User
In The Loop

Adaptive

Invisible MT

Interactive

Invisible MT

More Accurate

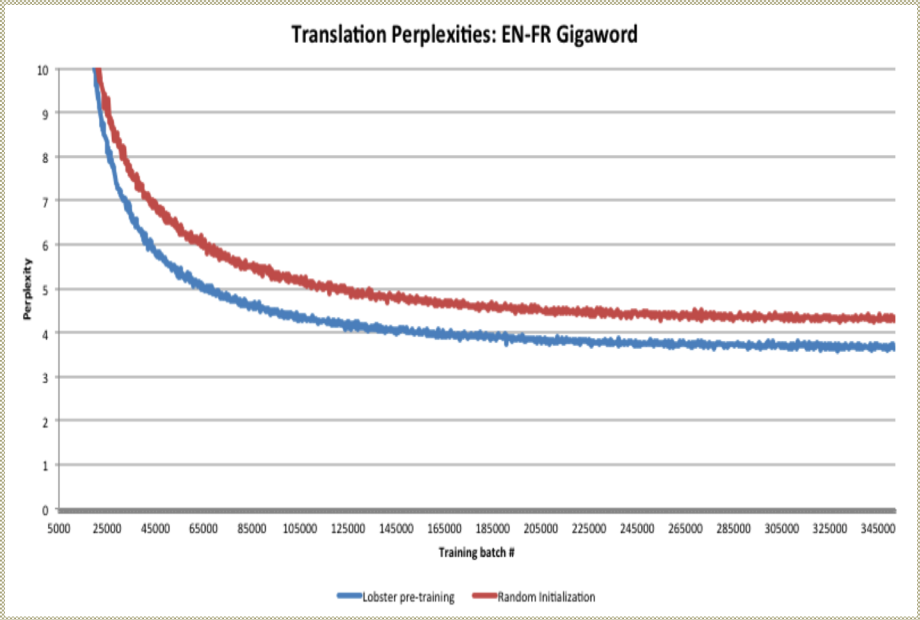
Work to date…

PROTOTYPE


Normal Layout her
Note
Remind me
Type below to test your settings
خليج فارس ٠١٢٣ همي
Compose
To
Subject
Bei
-John
AZERTY
dh
Henkel
שלח
anysoftkeyboard download
Paramètres de AnySoftKeyboard
New message
To:
Send
216 messages sent
To
Ше
шесть

Deep Learning for MT

- Train a recurrent neural network to understand input and generate translation
- These models run on the mobile device
- Tunable to specific domains
  - Leverages monolingual text data
  - Can be retrained



DEEP LEARNING FOR MT

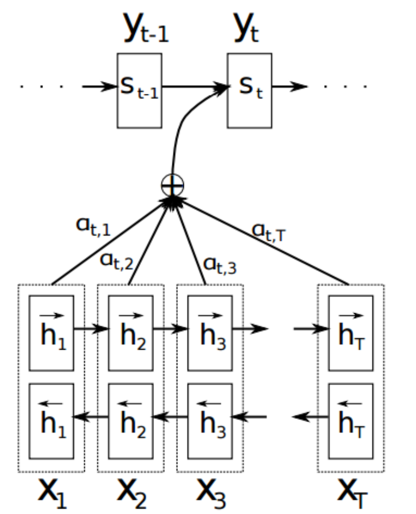


Figure 1: The graphical illustration of the proposed model trying to generate the t-th target word y_t given a source sentence $(x_1, x_2, \ldots, x_T)$.

Bahdanau, D., Cho, K. and Bengio, Y., 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

We're currently following the standard sequence-to-sequence model for MT.

One part of model learns where to look in the source sentence when trying to produce the next word. *Attention*

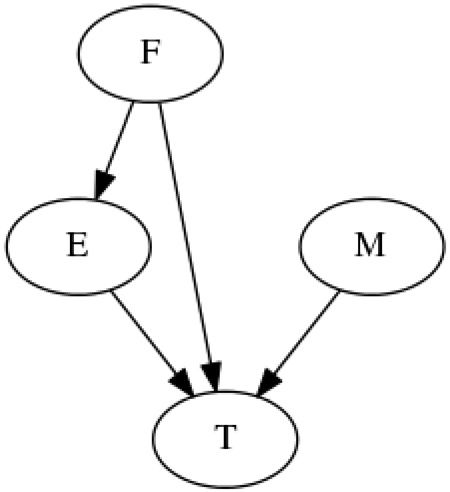
Other part of the model decides what word to produce given where the first part is looking.

THE GRAPHICAL MODEL

$$P(F,E,T,M) = \sum_{F,E,M} P(F)P(E|F)P(T|F,E,M)P(M)$$

$$= P(M=1)\sum_{F} P(T|F)P(F)$$

$$+ P(M=0)\sum_{E,F} P(T|E)P(F|E)P($$



$P(E)$ English Language Model

$P(F)$ Foreign Language Model

$P(F|E)$ Translation Model

$P(T|E), P(T|F)$ Editing Models

$P(M)$ Modality (1=typing Foreign, 0=typing English)

TRAINING & TESTING DATA

Europarl Data

Not the right fit – formal

Not the type of language used for texting - informal

Online Movie Database – OpenSubtitles

Good fit

Conversational

Pierre Lison and Jörg Tiedemann, 2016, *OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles.* In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016).

language	files	tokens	sentences	af	ar	bg	bn	br	bs	ca	cs	da	de	el	en	eo	es	et	eu	fa	fi	fr	g
af	32	0.2M	27.4k		6.2k	7.6k			1.8k		10.5k	6.0k	7.9k	11.6k	16.2k		12.6k	2.2k		2.1k	2.8k	7.3k	
ar	67,608	329.8M	60.8M	6.2k		16.2M	62.2k	13.0k	6.1M	0.3M	16.5M	7.5M	7.1M	15.3M	19.4M	19.4k	18.3M	6.9M	0.1M	3.0M	10.8M	14.4M	44
bg	90,376	523.4M	80.2M	7.7k	17.8M		60.7k	13.8k	7.5M	0.3M	21.1M	8.2M	8.9M	19.3M	26.4M	23.4k	24.8M	7.7M	0.1M	2.8M	13.2M	18.5M	48
bn	76	0.6M	0.1M		64.1k	62.7k			36.6k	3.1k	61.1k	58.2k	54.7k	58.5k	69.3k		65.8k	56.5k	3.1k	44.8k	56.1k	59.3k	
br	32	0.2M	23.1k		13.3k	14.1k			2.7k	5.3k	14.5k	10.0k	7.5k	14.4k	17.7k	1.1k	15.6k	15.0k	0.7k	4.4k	8.1k	15.4k	0
bs	30,511	179.5M	28.4M	1.8k	12.2M	8.5M	37.7k	2.7k		0.1M	7.5M	3.7M	3.6M	7.3M	9.5M	7.4k	9.0M	3.5M	76.3k	1.3M	5.2M	6.8M	27
ca	711	4.0M	0.5M		0.3M	0.3M	3.2k	5.5k	0.1M		0.3M	0.2M	0.2M	0.3M	0.4M		0.4M	0.2M		96.2k	0.2M	0.3M	11
cs	125,126	715.3M	112.8M	10.7k	18.1M	24.7M	63.3k	14.8k	8.5M	0.4M		8.5M	9.3M	19.8M	27.5M	31.7k	25.9M	7.9M	0.1M	2.9M	13.7M	19.1M	68
da	24,079	162.4M	23.6M	6.1k	8.0M	9.3M	60.9k	10.1k	4.0M	0.2M	9.6M		4.9M	8.1M	9.4M	11.3k	9.1M	5.0M	87.7k	2.1M	7.9M	7.6M	28
de	27,742	186.3M	26.9M	8.0k	7.6M	10.0M	57.2k	7.7k	4.0M	0.2M	10.6M	5.4M		9.1M	11.5M	24.9k	10.8M	4.3M	75.7k	1.8M	6.9M	9.2M	52
el	114,230	683.1M	101.6M	11.8k	16.8M	22.3M	60.5k	14.6k	8.1M	0.3M	23.0M	9.1M	10.2M		25.6M	24.5k	24.5M	7.5M	0.1M	2.8M	13.1M	19.6M	66
en	322,294	2.5G	336.6M	16.7k	21.9M	31.6M	75.0k	18.5k	11.1M	0.4M	33.8M	11.0M	13.4M	30.4M		49.0k	40.0M	8.6M	0.2M	3.3M	16.8M	28.0M	0.1
eo	89	0.5M	79.3k		19.9k	24.3k		1.1k	7.6k		32.8k	11.7k	25.6k	25.2k	51.1k		38.6k	17.6k		5.1k	18.9k	28.3k	0
es	191,987	1.3G	179.2M	12.9k	20.3M	29.2M	69.1k	16.0k	10.2M	0.4M	30.7M	10.4M	12.4M	28.6M	50.1M	40.2k		8.3M	0.2M	3.1M	15.7M	25.8M	0.2
et	23,515	140.7M	22.9M	2.2k	7.5M	8.9M	58.6k	15.4k	4.0M	0.2M	9.2M	5.7M	4.8M	8.6M	10.3M	18.2k	9.6M		93.3k	1.9M	6.5M	6.9M	29
eu	188	1.4M	0.2M		0.1M	0.1M	3.3k	0.7k	80.9k		0.1M	93.2k	80.1k	0.2M	0.2M		0.2M	0.1M		43.1k	0.1M	0.1M	10
fa	6,469	44.3M	7.4M	2.1k	3.1M	2.9M	46.3k	4.4k	1.4M	0.1M	3.1M	2.2M	1.9M	3.0M	3.6M	5.2k	3.3M	2.1M	44.7k		2.4M	2.5M	21
fi	44,594	208.5M	38.7M	2.8k	11.5M	14.8M	57.9k	8.3k	5.7M	0.2M	15.3M	9.0M	7.6M	14.6M	19.2M	19.5k	17.7M	7.4M	0.1M	2.5M		12.5M	40
fr	105,070	672.8M	90.9M	7.5k	15.5M	21.3M	61.4k	16.3k	7.4M	0.3M	21.8M	8.5M	10.3M	22.2M	33.5M	29.1k	30.1M	7.8M	0.1M	2.7M	13.9M		93
gl	370	1.9M	0.2M		45.8k	49.7k		0.5k	28.0k	11.9k	71.3k	29.4k	54.5k	68.8k	0.2M	0.3k	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	96.1k	

EXAMPLE TRANSLATION MODEL ENTRIES

almost	
presque	39%
près	19%
pratiquement	9%
quasi	5%
quasiment	4%

some	
certains	28%
certaines	11%
une	7%
quelques	6%
des	6%
un	5%

border	
frontière	24%
frontières	20%
des	10%
aux	10%
frontalières	6%
les	5%
frontaliers	5%
frontalière	3%
la	3%
frontalier	3%

youth	
jeunesse	25%
jeunes	22%
la	19%
des	13%
les	6%
pour	3%
de	2%

DEEP LEARNING FOR MT

Arabic-English BLEU

15.1	Baseline
18.7	+UNK, some post-processing
22.4	+gradient clipping, longer training
24.5	+pretrained word embeddings
…	

System is now near state-of-research performance

- Time to switch to E-A for MCT

DEEP LEARNING FOR MODEL COMPRESSION

The amount of storage on phones is smaller compared to that on servers

- Neural methods can be used to make small models as effective as traditional, larger models

T-TABLE COMPRESSION

Input: English word

"Feast":

{

"festin": 0.42, "fête": 0.57

}

Output: Probability it translates to each of 43K French words (most ~0%)

Challenge: Model table with fewer parameters than needed to store (table entries)

Score: Cross Entropy (CE)

- Lower means less precision lost during compression.

T-TABLE COMPRESSION, *WORK IN PROGRESS*

Source side: English word

Target side: French word

Numeric entry: $p(f|e)$

Compression of source side

5.3 CE Recurrent Neural Net (RNN)

0.75 CE Convolutional Neural Net (CNN)

Target side compression (all infeasible for use in an actual system)

0.73 CE Word Index

5.4 CE RNN

5.2 CE CNN

T-TABLE COMPRESSION, BOTTOM LINE

Can we reduce data storage requirements while still predicting MCT suggestions with enough accuracy to be useful?

- Yes

Best approach we've found, *so far:*

- Character-based input encoder, one character per keystroke typed
- Hierarchical softmax output decoder produces the distribution of numbers over the output vocabulary

Next Steps

1. Obtain user Feedback

2. Determine if data used designing the system is adequate

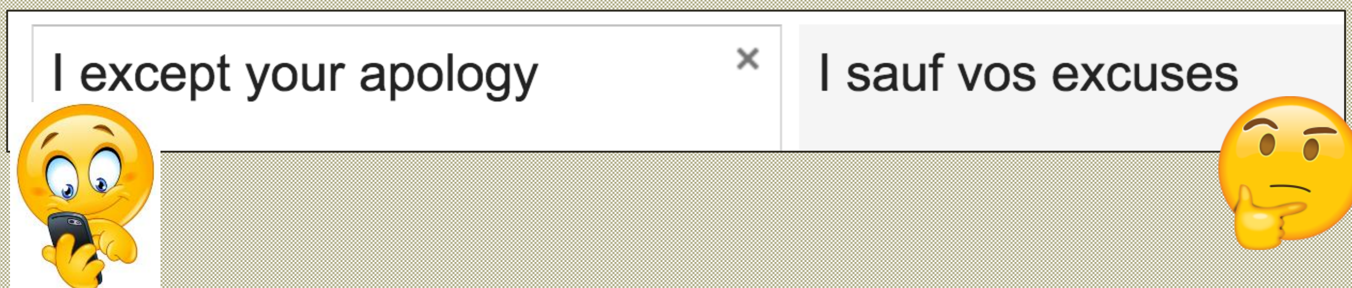
4. Look at if there are areas to optimize

Questions?

DEEP LEARNING FOR ERROR CHECKING AND HINTING

Rapid typing encourages certain types of mistakes

- Some, e.g. typos, can be corrected immediately
- Others require context: determiners, inflections, homonyms



Magic Punctuation: when the author types a period, can we identify if part of the sentence doesn't match the user's intent?

DEEP LEARNING FOR ERROR CHECKING: MANDARIN

Goal: identify when user selects the wrong character from a list of phonetically similar options with same pinyin transcription

Our noise model samples from a pinyin / character frequency table to corrupt clean Mandarin sentences

- Allows us to cheaply build large training data from monolingual text

Our denoiser is a recurrent neural network that reads the sequence of characters, then "points" at a position in the input

- Very compact model (3 MB)
- Corrupt characters are detected with over 85% accuracy in initial tests