Automatic Post-Editing and Machine Translation Quality Estimation at eBay

Nicola Ueffing
eBay MTScience Team
2018-03-21, AMTA Workshop

Intro

Nicola Ueffing

- Research scientist on eBay‘s machine translation research team since May 2016
  - machine translation for e-commerce content and for natural language generation (incl. APE)
  - A bit of quality estimation
- Prior to eBay:
  - research scientist at Nuance Communications (e.g. Dragon NaturallySpeaking)
  - PostDoc at Interactive Language Technologies team, National Research Council Canada
  - PhD in computer science from RWTH Aachen University: confidence estimation for machine translation

Overview

Why MT at eBay?

Automatic Post-Editing for Browse Page Titles

MT Quality Estimation for e-commerce content

Ongoing research

Why MT at eBay?

170M
active buyers

57%
of business is international

190
Markets

1.1B
live listings

Q4 2017

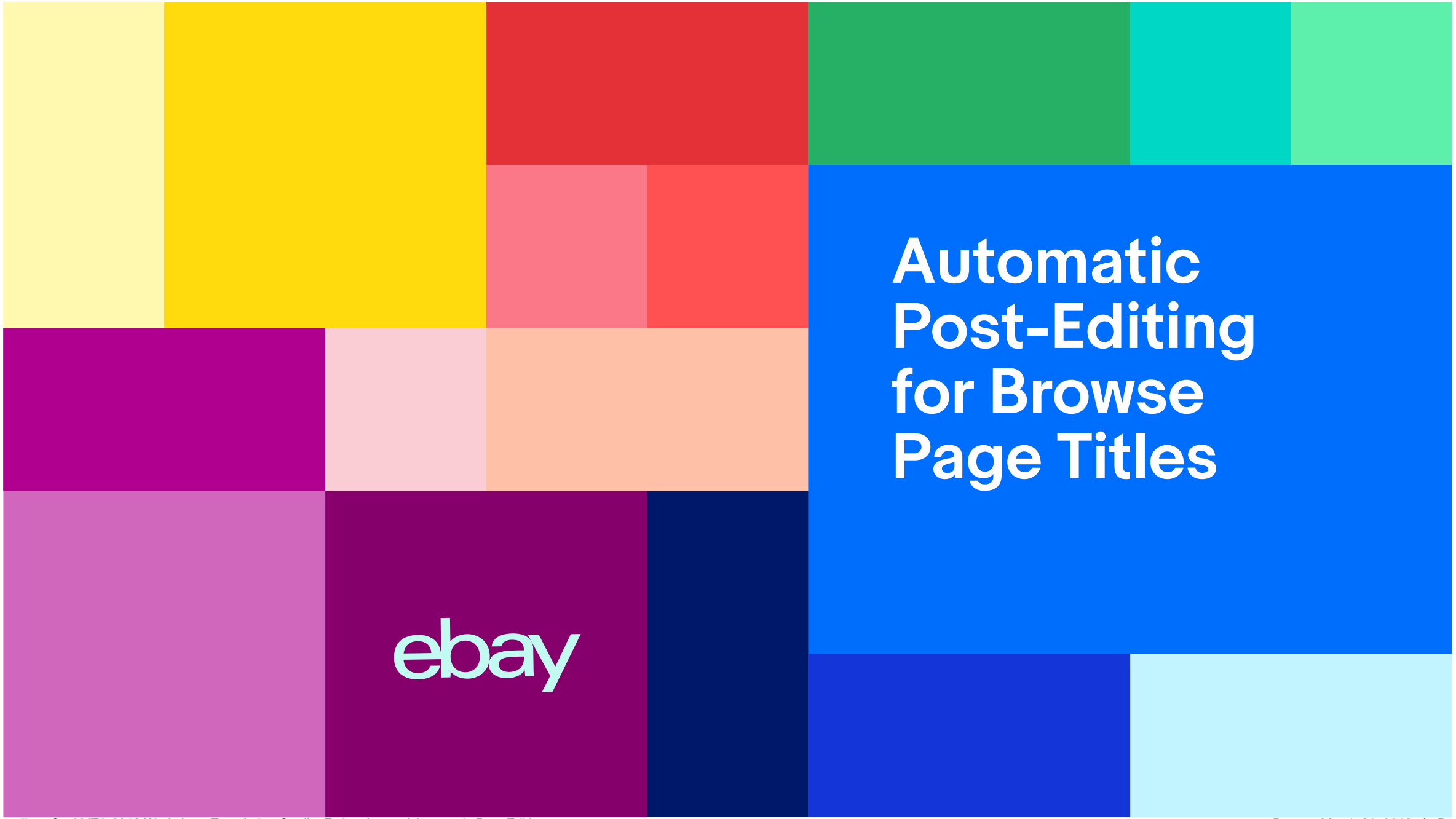
Applications of MT technology

Machine Translation

- Enable cross-border trade
- Translate
  - Search queries
  - Item titles
  - Item descriptions

Browse Pages: Title Generation

- Translate name-value pairs describing items into natural language


Automatic
Post-Editing
for Browse
Page Titles
ebay

How to explore the many items on eBay?

Browse Pages

Idea:
Create permanent “browse” pages for all items & products within a category that share a certain set of name-value pairs, e.g.

- In category “Light Bulbs”
- ”Wattage” = “9W”
- ”Bulb Shape Code” = “E27”

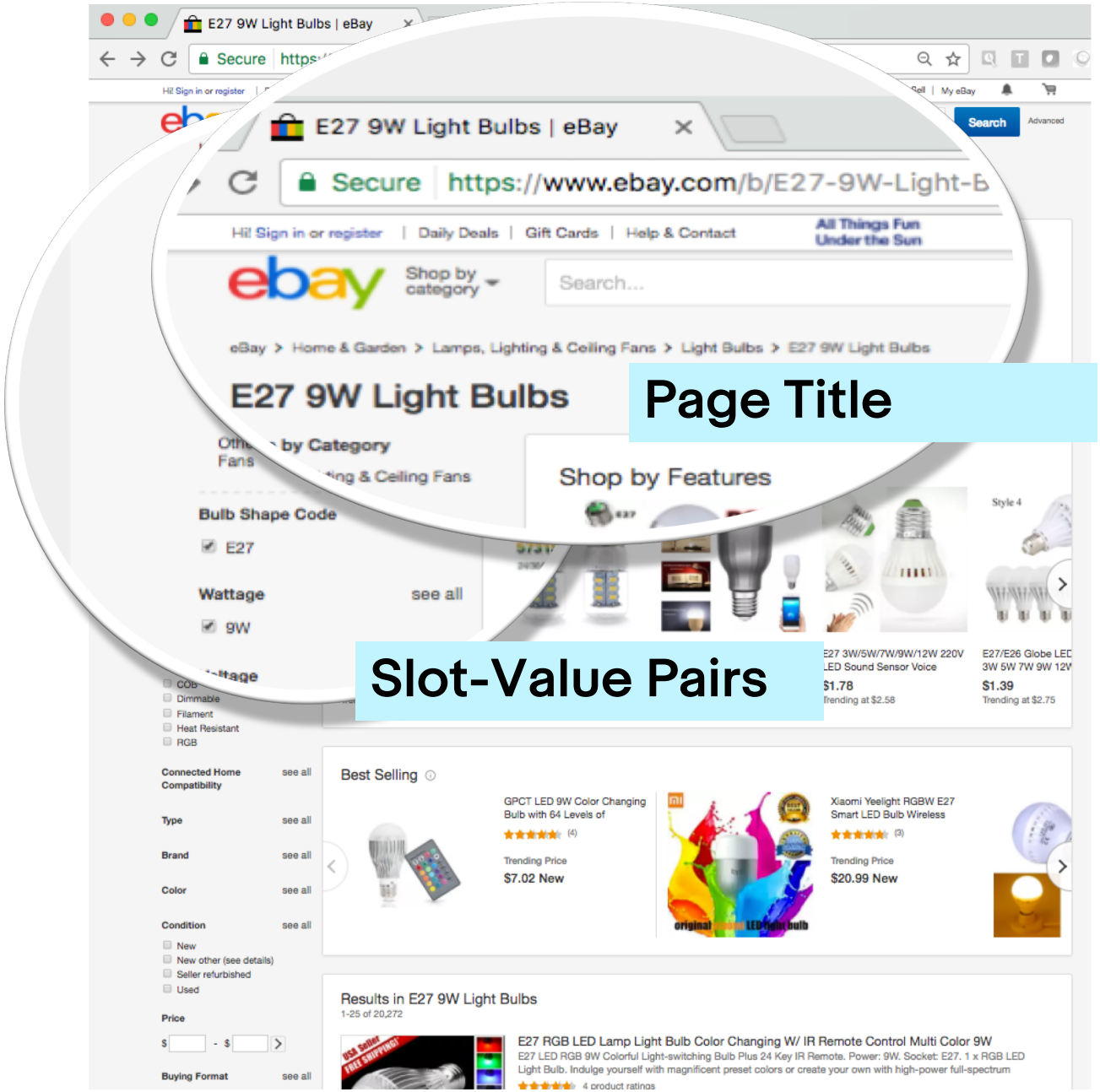
Users can then navigate to

- Related/refined browse pages
- Hot offers
- Individual products

=> Also beneficial for Search Engines

How to explore so many items?

Browse Pages



Why automatic title generation?

eBay is present

- in dozens of countries
- with thousands of categories
- with hundreds of thousands of name-value pairs (products aspects aka slots)

→Millions of potential browse pages (and titles) required!

Browse Pages

Step 1: rule-based title generation

Browse Pages

First approach we implemented for German:
Rule-based approach

1. Use hand-written heuristics / shallow parsers to classify each slot
2. Order slots based on slot classes
3. Realize each slot separately based on slot class
    - Use dedicated heuristics for certain combinations, e.g. Category + Product Type
4. Concatenate realizations

Step 2: APE

Browse Pages

For German, we have

- Millions of browse page titles in a slightly artificial language
  (our output from rule-based system)
- Parallel titles in a “natural” language (human curated titles)

=> train an APE system on those

e.g. translate
Kaukasische Wohnraum-Teppiche für Patchwork
into
*Kaukasische Wohnraum-Teppiche **mit** Patchwork-**Muster***

APE
Pros & Cons

Browse Pages

Pro

+ Straight forward
+ Large improvements in quality
+ Easy to integrate

Con

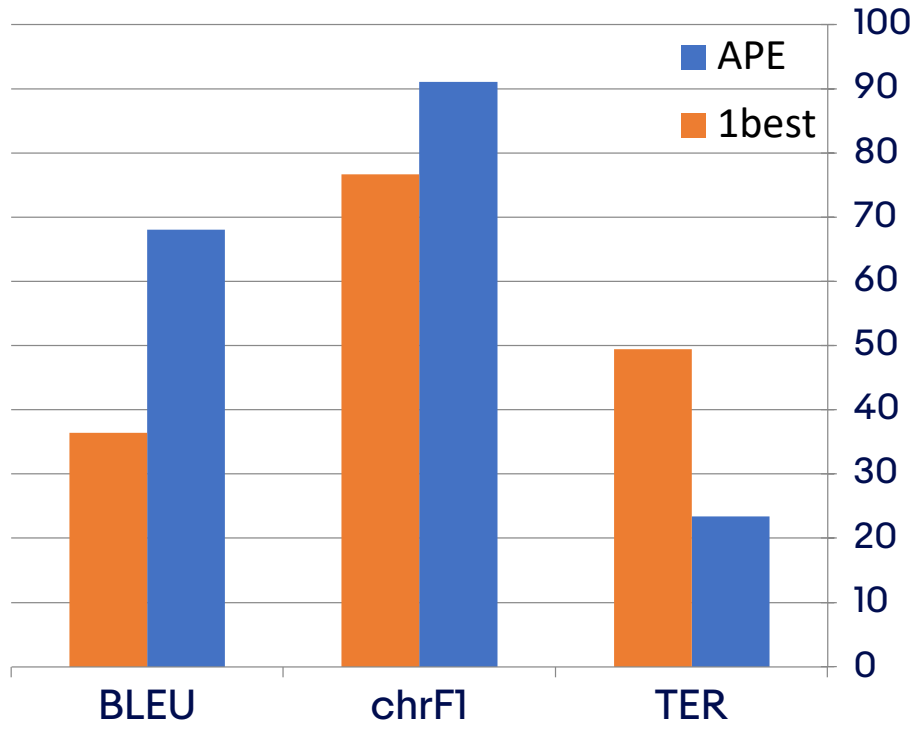
- Can only fix data that’s there (can’t reconstruct missing slots, slot names or context, ...)
- Sometimes learns artifacts from data (esp. when noisy)
- Will learn curation rules present when titles were created

APE Evaluation Results

Browse Pages

corpus	curated titles: #tokens
train	3.8M
dev	8.8k
test	8.8k

Evaluation on test



MTQE for e-commerce content

eBay item titles

Intro

Item Titles

- Relatively free word order
- +adequacy
- -fluency

Categories (e-commerce), e.g.

- Cellphones & Smart Phones
- Women's Clothing
- Car Parts & Accessories
- Cycling
- Fishing
- Skin Care
- Jewelry
- ...

eBay item titles

Intro

Examples:

- For Samsung Galaxy S5 i9600 S V TPU Crystal Clear Soft Case Ultra Thin Cover NEw
- 0.3mm Thin Crystal Clear Soft Silicone Fitted Case Skin Cover For iPhone 6 4.7"
- Universal 12000mAh Backup External Battery USB Power Bank Charger for Cell Phone
- Luxury Slim Aluminum Alloy Metal Bumper Frame Case/Cover For Apple iPhone 5 5S
- Luxury Ultra thin Metal Aluminum Bumper Case PC Cover For Samsung Galaxy Note 3
- 50000mAh Portable Super Solar Charger Dual USB External Battery Power Bank DX
- Sausage boiler broth boiler butcher's boiler boiler pot boiler insert
- Rasta wig with dreadlocks Rasta Hat Rasta braids
- CUTE HELLO KITTY Stuffed Plush 12" so CUUUUUUUTE!!!!(FREE SHIPPING in USA)

eBay item titles

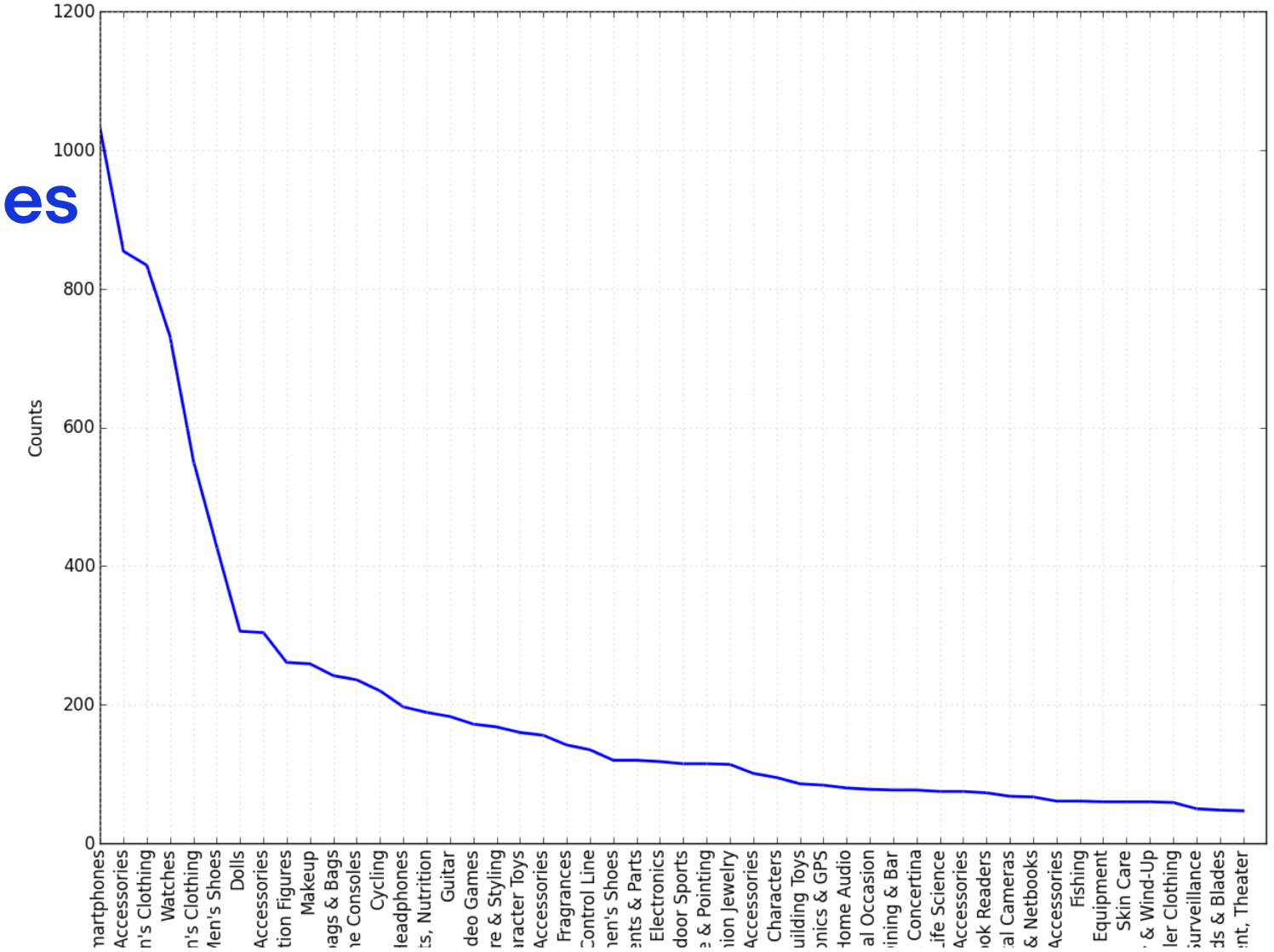
Data

- English-Portuguese
- Phrase-based Statistical MT
- Based on post-edition effort (HTER)
- Approx. 11k translated segments which are post-edited
- 223 different e-commerce categories

Data

eBay item titles

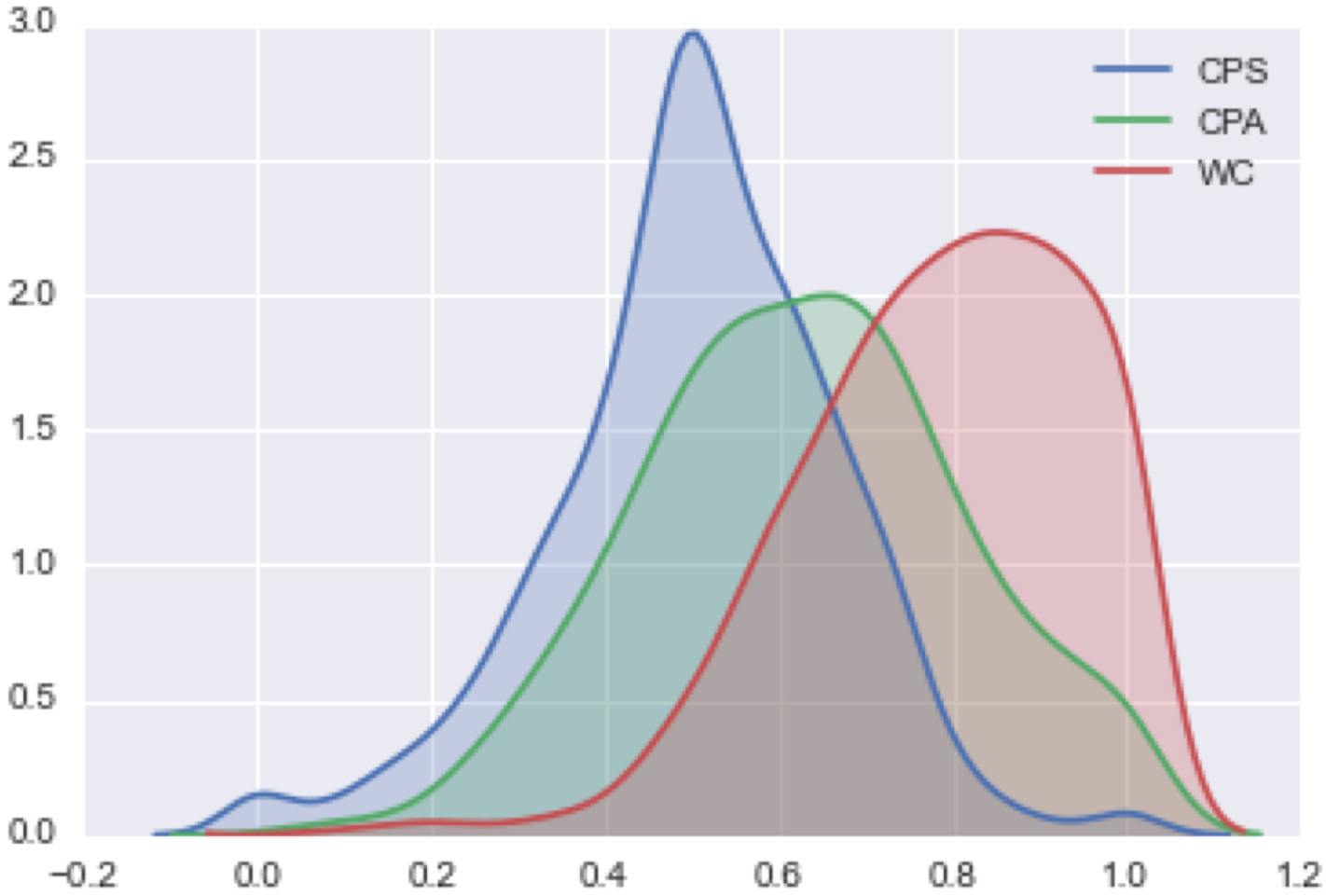
e-commerce categories



eBay item titles

Post-edition effort per category

Distribution of HTER for top 3 categories



Quality Estimation

Features

79 QuEst features:

- Black-box
- Complexity
- Adequacy
- Fluency

Item title embeddings

- Adequacy
- Concatenation of source and translation embeddings
- From paragraph2vec

NER-based

- Adequacy
- Numbers and ratio of NER tags found in source and translation

Quality Estimation

Learning algorithms

Extremely Randomized Trees

- Ensemble of decision trees
- Random forests
  - Build on random samples from training data
  - Choose best split for random subset of features
- Extremely randomized: additionally choose best threshold from random set of thresholds

AdaBoost

- Sequence of weak learners (very small decision trees)
- Fit them on original dataset
- Then fit additional copies of classifier on same data, but adjust weights of incorrectly classified instances s.t. subsequent classifiers focus more on difficult cases
- Final prediction: weighted majority vote of all iterations
- Time consuming

Both:

- Non-linear
- Provides feature importances

Quality Estimation

Experimental setup

- regression
- HTER labels clipped in [0, 1]
- 75/25 train/test splits
- Model selection
  - Randomized search with 5-fold cross validation (100 iterations)
  - Optimized for mean absolute error
- Evaluation
  - mean absolute error (MAE) ↓
  - Pearson‘s correlation ↑

Quality Estimation

Experimental results I

Cellphones & Accessories

	Extremely Randomized Trees		AdaBoost	
	MAE↓	Pearson↑	MAE↓	Pearson↑
Baseline: Mean	15.4	0	15.4	0
QuEst79	14.3	47.3	13.6	50.3
QuEst79 + embeddings	14.3	47.6	13.8	46.4
QuEst79 + NER	**13.8**	**50.4**	**13.1**	**56.0**
QuEst79 + NER + embeddings	**13.8**	49.9	13.5	51.9

Quality Estimation

Experimental results II

Cellphones & Smartphones

	Extremely Randomized Trees		AdaBoost	
	MAE↓	Pearson↑	MAE↓	Pearson↑
Baseline: Mean	12.9	0	12.9	0
QuEst79	12.4	39.6	11.7	45.6
QuEst79 + embeddings	12.5	38.7	12.2	41.6
QuEst79 + NER	**12.2**	**44.2**	**11.1**	**53.5**
QuEst79 + NER + embeddings	12.3	43.4	11.8	49.3

Quality Estimation

Experimental results III

Women's Clothing

	Extremely Randomized Trees		AdaBoost	
	MAE↓	Pearson↑	MAE↓	Pearson↑
Baseline: Mean	13.0	0	13.0	0
QuEst79	**12.8**	**13.2**	13.1	6.8
QuEst79 + embeddings	12.9	10.0	**12.6**	**11.3**
QuEst79 + NER	12.8	12.2	12.9	10.8
QuEst79 + NER + embeddings	12.9	7.2	12.7	4.1

- Fewer named entities than other 2 categories
- More generic description of items

⇒ NER not very helpful

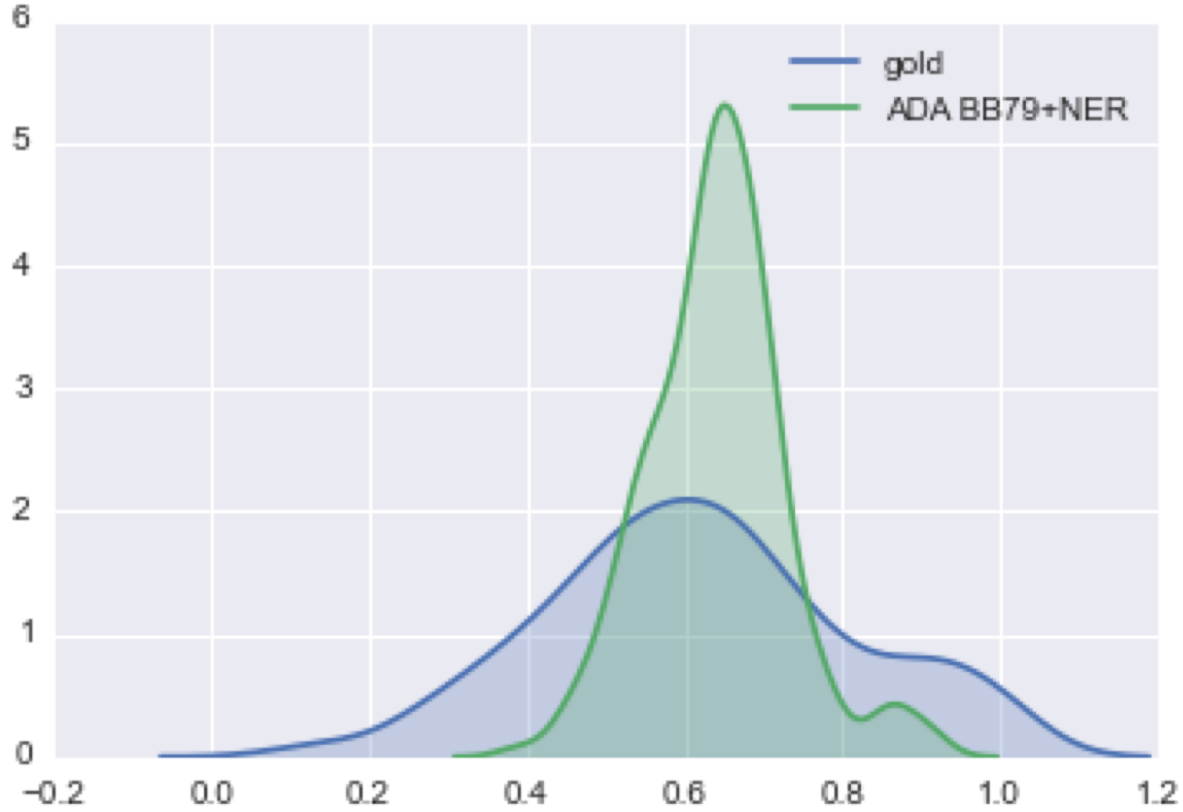
- Many bad translations

Quality Estimation

Analysis

Analysis

- Quality prediction in the tails of the test set distribution is problematic
- Tails equals to
  - Good translations (HTER close to 0)
  - Bad translations (HTER close to 1)



Quality Estimation

Analysis

Analysis

Best model, AdaBoost:

- Accuracy @ 25% worst translations (HTER near 1)
  - CPA: 52.83
  - CPS: 53.12
  - WC: 32.69
- Accuracy @ 25% best (HTER near 0)
  - CPA: 60.37
  - CPS: 43.75
  - WC: 30.76
- Random guess (baseline): ~25%

Quality Estimation

Conclusion

- Best feature set on average: Quest79 + NER
- AdaBoost presents the best accuracy, but slow
- Extremely Randomized Trees offer best trade-off between accuracy and computing time
- Models can predict bad and good translations with more than 50% accuracy
- Models for single categories, no pooling

Ongoing research

Ongoing research

Ongoing research

- User feedback from star ratings => bandit learning
- Quality estimation for natural language generation (browse page titles)
  - Random forest with features, mix of common and task-specific
  - Neural approach
- (Potential) QE applications
  - Do not display low-quality MT/NLG on site
  - Decide about updating existing title / translation
  - Routing for post edition
  - Data selection for post edition

References

Browse page title generation: APE approach and other MT-based methods described in:
International Conference on Natural Language Generation, Santiago de Compostela, Spain, September 2017
Generating titles for millions of browse pages on an e-Commerce site
Prashant Mathur, Nicola Ueffing, Gregor Leusch

Quality Estimation research described in:
MT Summit - User's Track, Miami, Florida, October 2015
MT Quality Estimation for E-Commerce Data
Jose G. C. de Souza, Marcello Federico, Hassan Sawaf

http://research.ebay.com/research-areas/research-machine-translation


Thank you
to my colleagues
José GC de Souza,
Prashant Mathur,
Gregor Leusch
ebay

Thank You