facebook

facebook

Extracting Translation Pairs from Social Network Content

Matthias Eck, Yury Zemlyanskiy, Joy Zhang, Alex Waibel
December 4th, 2014

Translations on Facebook

Over 1 billion people use Facebook

- Majority is located outside of US/English speaking countries
- Users generate content in many different languages

Monthly active users	
US & Canada	206 million
Europe	296 million
Asia	426 million
Rest of the world	423 million
Overall	**1.35 billion**

Translation is necessary to connect people across country and language barriers

- Cross country friendships
- International news and entertainment
- International celebrities

Goal & challenges

Goal:

Translate user generated content between languages

Challenges:

- Informal language: slang terms, internet language, …
- Spelling errors
- Social network terms: Facebook/Like/Share/Follow…
- No in-domain training data available

User content examples (public):

Guuysss !! Follow him! He made a new facebook! Like & Share

Good night fans, see ya.

Happpyyyy Birthdaaayyyy!!!

Lol nd once again she playing yo u my boy ! smh uu have bad luck-jah

Approach

Collect additional in-domain training data from Facebook content to improve machine translation performance

2 Methods investigated:

1. Identify multilingual posts and extract translation pairs
2. Collect monolingual post translations sharing same URL

Evaluation:

Evaluated on in domain test sets for
Spanish-English and Portuguese-English

Method 1: Multilingual posts

Multilingual posts:

- (some) celebrities, news, small business and other users post in multiple languages
- International fan bases, friends, local language diversity

General Approach:

- Identify posts with multiple languages
- Extract language segments
- Classify segments to ensure the segments are actual translations and not only code-switching

Multilingual posts – Language ID

Happy Birthday mi brother – Feliz cumple mi hermano

Steps:

1. Identify likely language for each word in post

	happy	birthday	mi	brother	feliz	cumple	mi	hermano
Language ID	en	en	es	en	es	es	es	es

2. First estimation of number of words in each language (stop if too imbalanced)

3. Smoothen language ID to eliminate incorrectly identified languages

	happy	birthday	mi	brother	feliz	cumple	mi	hermano
Language ID	en	en	es	en	es	es	es	es
Smoothen	en	en	en	en	es	es	es	es

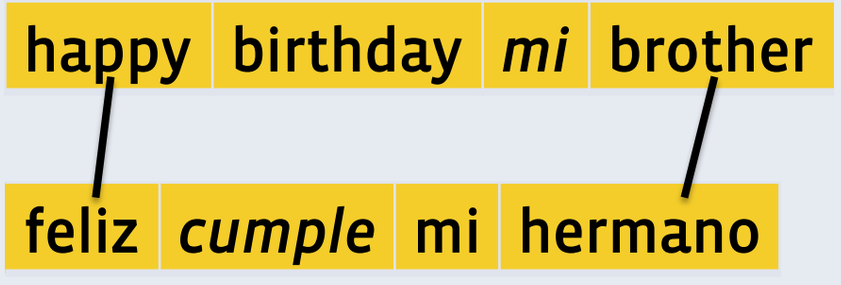
Multilingual posts – Segment and classify

Happy Birthday mi brother – Feliz cumple mi hermano

4. Segment post into language segments

happy	birthday	mi	brother	feliz	cumple	mi	hermano
en	en	en	en	es	es	es	es

5. Classify candidate pair as translation



Simple initial classifier:

#word to word translations found per segment length

cumple: previously unseen term for (informal) birthday, now covered as a translation

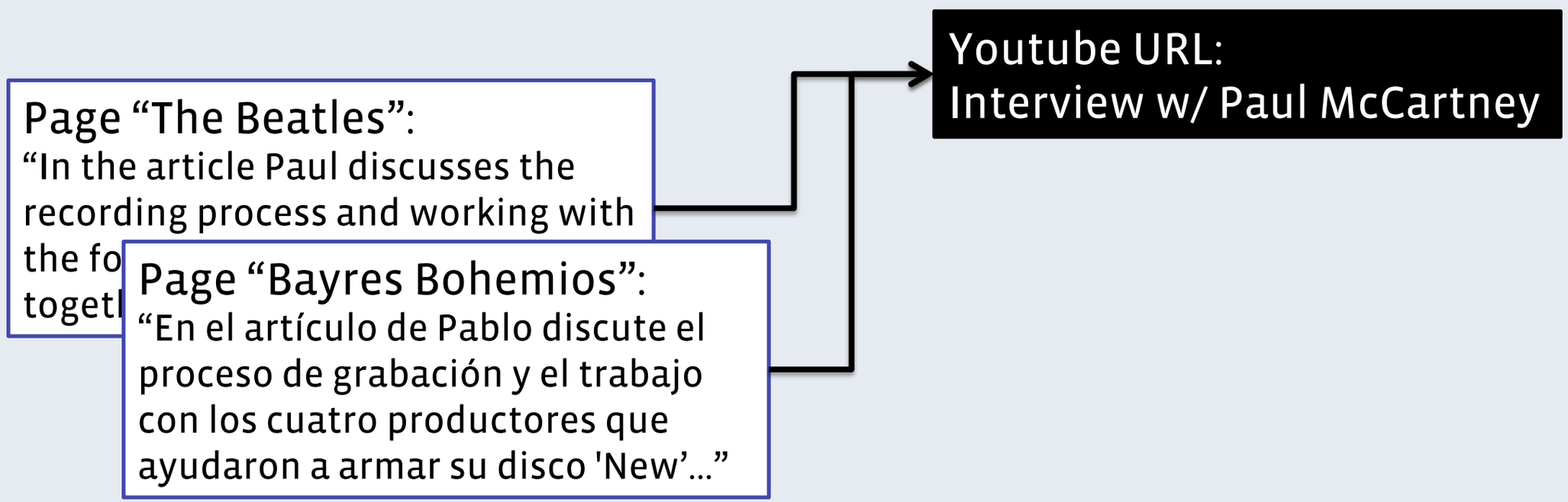
Method 2: URL shares

URL shares:

- Users can share web links (URLs) as part of their post
- E.g. quote from an article, mention the movie/song name etc. in different languages

General approach:

- Identify posts sharing same URL in different languages
- Classify to ensure posts are actual translations



Classifying candidate segments

Multilingual posts:

Simple classifier works well

- Either a post is translated as a multilingual post
  or
  code-switched and segments are not close

“quality time con mi chiqui”

URL shares:

Posts sharing the same URL can be very similar, but not translations

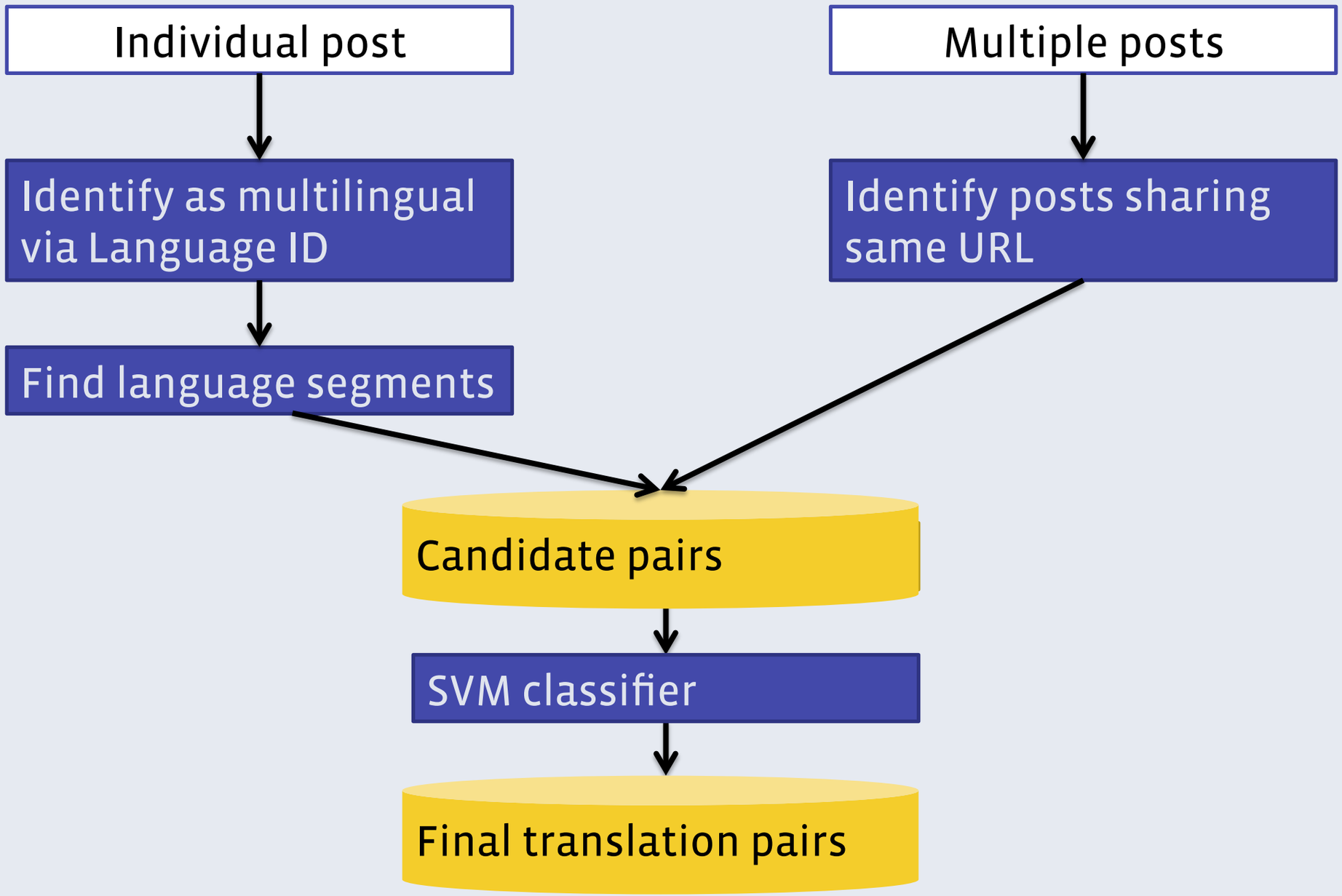
More advanced classifier is necessary to identify actual translations

Classifying candidate segments

SVM classifier with 25 features

- Length ratio
- All-to-all alignment features
  - Total IBM score
  - Maximum fertility
  - Number of covered words
  - Length of longest sequence of covered words
  - Length of longest sequence of not-covered words
- Max alignment features
  - Total IBM score
  - top 3 fertility values for target sentence
  - number of covered words for target sentence
  - “maximum intersection”: (maximal number of consequent source words, which have corresponding consequent target words)
  - maximum number of consequent not-covered words in target sentence

Process Overview



Experiments

Baseline Data

- Dev & Test data:
  2000 lines from public Facebook posts translated
  1000+1000 lines dev + test
- Training data:
  500,000 lines of EPPS (European parliament parallel data)
- Subset of full EPPS data – sorted according to estimated importance

Training procedure:

- Standard Moses system
- mgiza implementation of giza++
- 3-gram SRI language model with Kneser-Ney discounting
  trained on target side of parallel corpus
- Minimum Error rate training on dev set

Baseline results

Spanish-English

	Es-En	En-Es
Baseline Data (words)	*500,000 lines (8.48M/8.44M)*	*500,000 lines (9.29M/10.06M)*
Baseline BLEU	22.08	22.48
Baseline OOV rate	8.7%	12.9%

- Relatively high OOV rates, especially on English with many internet & slang terms

Portuguese-English

	Pt-En	En-Pt
Baseline Data (words)	*500,000 lines (11.29M/11.26M)*	*500,000 lines (11.26M/12.24M)*
Baseline BLEU	28.39	26.87
Baseline OOV rate	7.9%	10.8%

Collected Data

From Multilingual posts

	Spanish-English		Portuguese-English	
Multiling. posts	17,214 lines	925k/925k words	6,208 lines	241k/236k words

From URL shares:

- URL shares: 25 million & 9 million candidate pairs were found for Spanish & Portuguese

	Spanish-English		Portuguese-English	
URL shares	120,594 lines	2.91M/2.73M	95,444 lines	2.35M/2.28M

- Data collection is not directional and the resulting data was added to both translation directions

Results Spanish - English

BLEU scores and OOV rates:

	Es->En		En->Es	
	BLEU	OOV rate	BLEU	OOV rate
Baseline	22.08	8.7%	22.48	12.9%
+multi	23.47	7.8%	22.72	12.0%
+shares	23.16	6.0%	27.61	10.4%
+multi+shares	**24.30**	**5.9%**	**27.78**	**10.2%**

- For Spanish to English: Translations from multilingual posts have more impact
- For English to Spanish: Translations from URL shares have far more impact
- In both cases: Improvements are complementary

Results Portuguese - English

BLEU score and OOV rates

	Pt->En		En->Pt	
	BLEU	OOV rate	BLEU	OOV rate
Baseline	28.39	7.9%	26.87	10.8%
+multi	28.92	7.6%	26.95	10.5%
+shares	31.34	6.9%	31.11	9.1%
+multi+shares	**31.67**	**6.8%**	**30.92**	**9.0%**

- Here for both directions:
  Far better improvements with the data from URL shares
- Only small improvements using data collected from multilingual posts

Example translations & Analysis

Better coverage for OOV words

- OOV rate for Spanish to English dropped from 8.7% to 5.9%
- cargador and agrego

Note:
Spanish uses “like” directly

Examples: Spanish to English

Source	con el cargador incluido.
Baseline	with the *cargador* included.
Improved	with the charger included.
Reference	charger included.
Source	like y agrego !!
Baseline	like and *agrego*!!
Improved	like and add!!
Reference	like and add!!

Example translations & Analysis

Improved word usage:

- cumple: commonly used as “birthday” in “feliz cumple”

Improved phrase coverage/LM

- “memory card” instead of “card by heart”

Examples: Spanish to English

Source	feliz cumple preciosa !
Baseline	happy meets beautiful
Improved	happy birthday beautiful!
Reference	happy birthday, honey!
Source	sin tarjeta de memoria .
Baseline	without card by heart
Improved	without memory card
Reference	without memory card

Conclusions

Applied two methods to collect additional data from Facebook posts

- Multilingual posts (self translation in the same post)
- Monolingual posts sharing the same URL

Both methods enabled us to collect additional in-domain data

- Improvements are complementary and add up to combined higher scores
- Up to 5.2 BLEU improvements
- Significant drops in OOV rate (up to 30% relative improvements)
- Enhanced vocabulary & phrase coverage

facebook