Extracting Translation Pairs from Social Network Content

Matthias Eck, Yury Zemlyanskiy, Joy Zhang, Alex Waibel

Facebook, Inc.
eck@fb.com, urikz@fb.com, joyzhang@fb.com, waibel@fb.com

Abstract

We introduce two methods to collect additional training data for statistical machine translation systems from public social network content. The first method identifies multilingual content where the author self-translated their own post to reach additional friends, fans or customers. Once identified, we can split the post in the language segments and extract translation pairs from this content. The second methods considers web links (URLs) that users add as part of their post to point the reader to a video, article or website. If the same URL is shared from different language users, there is a chance they might give the same comment in their respective language. We use a support vector machine (SVM) as a classifier to identify true translations from all candidate pairs. We collected additional translation pairs using both methods for the language pairs Spanish-English and Portuguese-English. Testing the collected data as additional training data for statistical machine translations on in-domain test sets resulted in very significant improvements of up to 5 BLEU.

1. Introduction

Current social networking websites like Facebook, Twitter and LinkedIn are operating globally. The majority of Facebook's over 1 billion users[1] are located outside of the US and user generated content is produced in a wide variety of languages. A globalized world also supports friendships across country and language barriers and makes news and entertainment sources in other languages easily accessible. It is Facebook's stated mission to make the world more open and connected and giving people the power to share.
All of these facts generate the need for translation of user content. Efficiency and especially the amount of content requested to translate make only automatic translation systems feasible.

One of the main challenges in training translation systems for social media content is the lack of in-domain training data. Bilingual corpora are generally only available in news or parliament domains, which are considerably different from the actual content that needs to be translated in social media applications.
Social media content frequently exhibits slang terms, colloquial expressions and other features not common in carefully edited news sources. Spelling errors are also very frequent. Social media content in Spanish and Portuguese specifically often exhibits a lack of correct diacritical marks.

A general approach to overcome any domain-mismatch problem is to somehow collect additional in-domain training data to augment the out-of-domain training data. Many experiments could show that this often significantly improves the translation performance.

The source that is used here is the actual social network. This paper introduces two different approaches to automatically collect parallel training data from social network content.

1.1. Multilingual Posts

Posting the same content in many languages is an approach that many fan pages, but also individual persons take to reach different groups of their friends and fan bases. Popular fan pages on Facebook have up to 100 million and more fans. As of August 2014 e.g. singer Shakira has 102 million fans, soccer club FC Barcelona has 72 million fans and soccer player Lionel Messi has 69 million. All three are examples of fan pages that post most of their updates in English and Spanish (also Catalan in FC Barcelona's case). Figure 1 shows an example post by Lionel Messi.

Figure 1: Multilingual post by Lionel Messi in Spanish and English

> Gracias a mis compañeros por elegirme como uno de los capitanes del equipo y por la confianza que han depositado en mí. Un abrazo.
>
> Thanks to my teammates for picking me as one of the club captains and for the confidence they have given me. A hug.

These are just some of the millions of pages on Facebook. It is likely that many of them have a multilingual group of people following the page. In order to serve these people better a large number have resorted to multilingual posts. This is even the case for pages of smaller, local businesses. Many cities and communities in the United States for example have large ethnic minority populations, most notably people of Hispanic and Asian descent. To reach these potential

[1] Facebook has 1.35B monthly active users as of Sept. 30th, 2014 (Q3 2014 earnings call)

customers even small businesses often resolve to multilingual communication. These pages and users want to ensure that all language groups of their fans are appropriately informed without relying on machine translation, which might not be available on all platforms.
Our first approach determines if an individual post is part of this category and contains more than one language. Should this be the case the post is split into the individual language segments and a classifier decides if the parts are indeed translations of each other.

1.2. URL Sharing

The second approach exploits the sharing function in Facebook allowing users to publicly share and re-share links to videos or other websites. Users on Facebook and other social networks use this function to point their friends and colleagues to interesting content and can also comment on it separately. Popular videos, articles and websites are shared many times even across different language users.
The assumption here is that two users or pages talking about the same content might have very similar comments. Therefore we can consider the respective posts *comparable* and we try to find true parallel sentences among them. It is for example rather common for users to translate movie titles or to quote important parts of a news article in their own languages.
Recently, the official "The Beatles" page shared a YouTube video featuring Paul McCartney and wrote a description about it. The hotel "Bayres Bohemios" in Argentina then decided to share the video with its guests. They posted the same link with the same description translated to Spanish (see Figure 2)

Figure 2: Descriptions by the pages "The Beatles" and "Bayres Bohemios" for the same URL

URL:
https://www.youtube.com/watch?v=pE_1V0phMW8

"The Beatles":
Paul is interviewed in this week's NME Magazine, which is on the stands from today.
In the article Paul discusses the recording process and working with the four producers who helped put together his 'New' album; Paul Epworth, Ethan Johns, Giles Martin and Mark Ronson. The article reveals the name of two of the tracks from the album; 'Alligator' and 'Save Us'.

"Bayres Bohemios":
Paul es entrevistado en la revista NME de esta semana, que está en las gradas de hoy.
En el artículo de Pablo discute el proceso de grabación y el trabajo con los cuatro productores que ayudaron a armar su disco 'New', Paul Epworth, Ethan Johns, Giles Martin y Mark Ronson. El artículo revela el nombre de dos de las canciones del álbum, 'Alligator' y 'Save Us'.

The rest of the paper will discuss some related work in section 2 and describe our methods in sections 3 and 4. Sections 5 and 1 describe the data we were able to collect and our experimental results using this data to improve machine translation systems for Spanish-English and Portuguese-English.

2. Related work

Collecting corpora for machine translation is a well-researched problem. Collecting additional parallel sentences from Wikipedia and the web itself has been extensively studied due to the ease of access. [1]–[5]. Most approaches consist of two steps, identifying comparable candidate segment pairs based on some connection feature between them and a final step to classify the found candidate segments into actual translation pairs. A classification approach similar to [6] is generally applied. The importance of the accuracy of the classification is generally closely related on the method used to identify candidate segments.
Closely related to our multi-lingual post approach is the work done in [7] to collect additional Chinese-English translation pairs from Sina Weibo content. The authors continue the work in [8] by using crowdsourcing to improve the accuracy of the extracted data.

3. Collecting from multilingual Facebook posts

For all discussed experiments, only public posts were considered and in all instances these public posts were stripped of specific user attribution.

We generally consider all (public) Facebook posts as candidates for multilingual posts. At creation time of every Facebook post, a standard language identification system is applied. This helps with News Feed ranking and later the ability to show appropriate automatic translations.
Our translation extraction approach is now focusing on one source and target language pair at a time and we consider all posts that were identified as either target or source language in this step. The standard language identification does not consider multilingual posts and will only assign a single language identifier.

3.1. Language identification and segmentation

To identify the segments, we first apply an additional language identification step and decide for each unigram what its most likely language is.
Once the basic language identification is applied we also check if the ratio of terms identified as either language is within a reasonable range, otherwise the post is already discarded as unlikely to contain translated segments e.g. a post that contains ten English words and only one Spanish word.

In a second language identification step we apply a smoothing on the identified languages to eliminate spurious incorrect identifications. This changes the identified language of a single word if the neighboring words were identified as the other language. This has proven helpful for misspellings. Table 1 shows an example for a misspelling "mi" in the English segment. This is initially incorrectly identified as Spanish and then fixed in the smoothing step

Table 1: Language ID with smoothing

	Happy	birthday	mi	brother	…
Language ID	en	en	es	en	…
Smoothed	en	en	en	en	…

Once the language of every word has been identified the post is split into the two longest segments, which are then classified to determine if they are actually translations of each other.

3.2. Classifying the translation

All translation classifiers that were applied in this work are based on seed lexicons taken from the baseline trainings for each translation direction. This especially provides word-to-word lexicons to the classifiers.

Experiments have shown that in the multilingual post case even simple word-to-word translation heuristics provide adequate performance to distinguish candidates that are translations from ones that are not. The reason seems to be that in this case the users either actually provide a translation or they code-switched in their posts. In this case the segment contents are not close. An example post for this is "*quality time con mi chiqui"*[sic]. In this case there is little danger that the two segments could be classified as translations since no part of the segments are translations or even semantically close.

It is obviously also possible to apply more sophisticated segment classification and we describe a detailed model in section 4.2 originally developed to classify candidates generated from URL shares where candidates can often be much closer. The actual experiments reported all used the classifier described in section 4.2.

4. Collecting translations from URL Shares

An alternative idea to extract translations from Facebook posts is to try to find monolingual posts that are translations of each other. Of course it is not practical or reasonable to compare every post with every other post, so the idea is to preselect post pairs that are comparable, i.e. discuss the same content.
Our idea was to look at URL shares. Users in Facebook (and other social networks) have the ability to post links to web content outside of the social network. Should two users link to the same URL they are obviously commenting on the same content and it is likely that some of those users comments could be translations of each other.
Some examples are translated quotes from a news article, translated song, movie or book titles or just general comments like "*Great game by Germany in the world cup*". Given the vast number of users on popular social networks it is likely that a small number of them will then be actual translations that can be collected.

4.1. Collecting URL shares

As stated, the task of searching for parallel sentences in all possible combinations of monolingual posts is intractable. In addition to considering only monolingual posts in different languages, which shared the same URL, we also used a couple of other simple heuristics to further reduce the search space.
We split each post into individual sentences and compare all sentences in one language with sentences in other languages using these simple rules:

- Original posts share the same URL
- At most a length ratio of 2
- Difference between posts' creation times is no more than 3 days
- Three sequential words in one sentence translate with high lexical probability into three other sequential words in the other sentence.

These procedures can be efficiently performed in a MapReduce framework handling an enormous amount of data.
If we find a match between sentence A from post A^* and sentence B from post B^* we mark all possible pairs from A^* and B^* as candidates. This algorithm does not take the translation direction into account, so it has to be performed once per language pair.
Overall we identified 25 million candidate pairs for Portuguese-English and 9 million for Spanish-English (in the chosen timeframe).

4.2. Translation classifier

The final step is to filter parallel sentences from the prepared candidate pairs. It has been shown (in [9]–[11]) that SVM-based classifiers with lexical features are performing quite well for this purpose.
We rely on a combination of 25 features selected from [9]–[11]:

- ratio of number of words per sentence
- all-to-all alignment features (per each direction)
  - total IBM score (with all-to-all alignment)
  - maximum fertility
  - number of covered words
  - length of longest sequence of covered words
  - length of longest sequence of not-covered words;

Also all features except the IBM score are normalized by source sentence length.

- max alignment (per each direction)
  - 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 uncovered words in target sentence

Here all features are normalized by target sentence length except the IBM score (which is not normalized) and maximal intersection (which is normalized by source sentence length).

We used the same parallel corpora from the baseline machine translation training and tuned the classifier in order to achieve 95%-98% precision on the dataset. A possible problem here is that the data and users posts are essentially in different domains and the classifier might perform worse on our candidate pairs. It is common practice in this case ([11]) to run the filtering iteratively – using updated lexical dictionaries every time. However, it appeared to not be required, as the extracted corpora from the first iteration already gave a significant boost in translation quality.

The results show that the classifier filtered out 99% of the candidate pairs, but the remaining 1% was of very good quality – we did not find any non-parallel sentences while inspecting. The most common error was a few extra words in one of the sentences. The results show, that this does not negatively affect the final performance. Word and phrase extraction is generally robust if this does not occur too frequently.

5. Data Collection Statistics

Data for both methods was collected from public Facebook posts. The collected data is not directional and we used the data sets for tests in both directions. Table 2 shows the exact statistics for the collected data.

Table 2: Collected data statistics

	Es-En	Pt-En
Baseline data	*500,000 lines* 8.48M/8.44M Es to En 9.29M/10.06M En to Es	*500,000 lines* 11.29M/11.26M Pt to En 11.26M/12.24M En to Pt
Multilingual posts	17,214 lines 925k Es words 925k En words	6,208 lines 241k Pt words 236k En words
URL shares	120,594 lines 2.91M Es words 2.73M En words	95,444 lines 2.35M Pt words 2.28M En words

Spanish is more common on Facebook than Portuguese, which explains why more data could be collected for Spanish-English compared to Portuguese-English.

6. Translation Experiments

The developed methods were tested on two language pairs, Spanish-English and Portuguese-English for both translation directions each.

6.1. Training and Testing Data

For both language pairs development and test sets were created from manually translated public Facebook posts. Approximately 2,000 lines were translated and split into development and test sets.

The selected posts had previously been requested for automatic translation for the respective language pair, so they are exactly in-domain for the task and exhibit all the typical features.

The training data consists of out-of-domain data taken from European Parliament data (EPPS) and general phrases from the Tatoeba corpus[1]. The training data was sorted according to estimated importance [12] and only the top 500k sentence pairs were included in the training. The results showed that this did not result in any significant drop in translation performance and allowed for much faster training runs.

6.2. Machine Translation System

We used the open-source Moses statistical machine translation system [13]. All systems were trained following the standard training method using the parallelized implementation mgiza of giza++ [14], [15] and standard phrase extraction. The language models were regular 3-gram models with Kneser-Ney discounting. They were trained on the target side of the training data using the SRI toolkit [16], [17]. We applied standard minimum error rate training on our development sets and tested the systems on the separate test sets. All systems were evaluated using the standard BLEU metric [18].

6.3. Experimental Results

The experimental results in Table 3 illustrate the improvements for all four translation directions. Starting from the baseline scores we see varying improvements of up to 5.2 BLEU when using either approach. Even though the URL shares collected significantly more data, the multilingual post approach also results in significant BLEU improvements and it outperforms the approach for Spanish to English.

Combining both data sources generally further improves the performance, which indicates that the data collected is considerably different from each other. Inspection of the data confirmed this and it appears that the data from multilingual posts often contains sales offers and local events while the data collected from URL shares covers more popular culture, entertainment and politics.

[1] http://tatoeba.org

We also calculated the (token) out-of-vocabulary (OOV) rates for each dataset and this further explains the improvements. In every case the added data significantly improves the OOV situation. This is due to improved coverage of spelling errors, slang terms and Internet lingo.

The results also show that the URL shares approach generally gives greater improvements than the multilingual post extraction (with the exception of Spanish to English). The data extracted from multilingual posts does especially not perform very well for translations from English to Spanish or Portuguese, while it performs better for translations into English.

Table 3: Experimental Results – BLEU (token OOV rate in parentheses)

	Es→En	En→Es
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%)
	Pt→En	En→Pt
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%)

6.4. Example translations

In addition to the standard automatic BLEU metric we also analyzed how the additional data actually improved our translation systems by comparing baseline and improved translations. Table 4 shows some example translations from the Spanish to English translation system with the source and reference translations.

The first translation is a typical example of a concept "memory card" that is unlikely to be present in the out-of-domain data.

The second example illustrates an out-of-vocabulary term "agrego", which is not present in the baseline system and is then covered in the improved system. It also shows that the term "like" is directly used in Spanish instead of a Spanish term.

The next example shows how the translation of the Spanish term "cumple" is changed from the incorrect "meets" and the last example again contains a regular OOV term "cargador" that is not covered previously.

Table 4: Experimental Results - Example translations

Source	sin tarjeta de memoria .
Baseline	without card by heart
Improved	without memory card
Reference	without memory card
Source	like y agrego !!
Baseline	like and *agrego*!!
Improved	like and add!!
Reference	like and add!!
Source	feliz cumple preciosa !
Baseline	happy meets beautiful
Improved	happy birthday beautiful!
Reference	happy birthday, honey!
Source	con el cargador incluido.
Baseline	with the *cargador* included.
Improved	with the charger included.
Reference	charger included.

7. Conclusion

We presented two methods to collect additional translation pairs from public social network content, specifically public Facebook posts. First, we identified multilingual posts, where the actual posts contain their own translation. We also investigate extraction from "comparable" public posts identified by sharing the same URL.

Using both methods we are able to collect significant additional bilingual training data for the language pairs Spanish-English and Portuguese-English. Adding the collected data from either method to the overall training data improves the translation performance significantly with overall improvements of up to 5.2 BLEU. The main improvements are caused by enhanced vocabulary and phrase coverage of social network content. Both methods appear to collect data in slightly different topics and style, so the improvements are complementary and add up to combined higher scores.

Collecting translations based on the URL shares approach has the additional advantage to not be limited by language pairs that have a lot of need for multilingual posts and bilingual speakers; instead it can be more generally applied to any language pair.

8. References

[1] P. Resnik and N. A. Smith, "The Web as a Parallel Corpus," *Journal of Computational Linguistics*, vol. 29. pp. 349–380, 2003.

[2] Y. Zhang, K. Wu, J. Gao, and P. Vines, "Automatic Acquisition of Chinese-English Parallel Corpus from the Web," in *Proceedings of the 28th European Conference on Advances in Information Retrieval (ECIR 2006)*, 2006.

[3] K. Fukushima, K. Taura, and T. Chikayama, "A Fast and Accurate Method for Detecting English-Japanese Parallel Texts," in *Proceedings of the Workshop on Multilingual Language Resources and Interoperability (MLRI 2006)*, 2006.

[4] J. Uszkoreit, J. M. Ponte, A. C. Popat, and M. Dubiner, "Large Scale Parallel Document Mining for Machine Translation," in *Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010)*, 2010.

[5] F. Ture and J. Lin, "Why Not Grab a Free Lunch?: Mining Large Corpora for Parallel Sentences to Improve Translation Modeling," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT 2012)*, 2012.

[6] D. S. Munteanu and D. Marcu, "Improving Machine Translation Performance by Exploiting Non-Parallel Corpora," *Journal of Computational Linguistics*, vol. 31, no. 4. MIT Press, Cambridge, MA, USA, pp. 477–504, Dec-2005.

[7] W. Ling, G. Xiang, C. Dyer, A. Black, and I. Trancoso, "Microblogs as Parallel Corpora," in *Proceedings of the 51st Annual Meeting on Association for Computational Linguistics (ACL 2013)*, 2013.

[8] W. Ling, L. Marujo, C. Dyer, A. Black, and I. Trancoso, "Crowdsourcing High-Quality Parallel Data Extraction from Twitter," in *Proceedings of the Ninth Workshop on Statistical Machine Translation (WMT 2014)*, 2014.

[9] T. Herrmann, M. Mediani, J. Niehues, and A. Waibel, "The Karlsruhe Institute of Technology Translation Systems for the WMT 2011," in *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, 2011.

[10] J. R. Smith, C. Quirk, and K. Toutanova, "Extracting parallel sentences from comparable corpora using document level alignment," in *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT 2010)*, 2010.

[11] S. Hewavitharana, "Detecting Translational Equivalences in Comparable Corpora," 2012.

[12] M. Eck, S. Vogel, and A. Waibel, "Low Cost Portability for Statistical Machine Translation based on N-gram Frequency and TF-IDF," in *International Workshop on Spoken Language Translation (IWSLT 2005)*, 2005.

[13] P. Koehn, H. Hoang, and A. Birch, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, 2007.

[14] Q. Gao and S. Vogel, "Parallel Implementations of Word Alignment Tool," in *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, 2008, pp. 49–57.

[15] F. J. Och and H. Ney, "A Systematic Comparison of Various Statistical Alignment Models," *Journal of Computational Linguistics*, vol. 29. pp. 19–51, 2003.

[16] A. Stolcke, "Srilm — an Extensible Language Modeling Toolkit," *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2002)*, vol. 2. 2002.

[17] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "SRILM at Sixteen : Update and Outlook," in *Automatic Speech Recognition and Understanding Workshop (ASRU 2011)*, 2011.

[18] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*, 2002.