Translating Government Agencies' Tweet Feeds: Specificities, Problems and (a few) Solutions

Fabrizio Gotti, Philippe Langlais
{gottif,felipe}@iro.umontreal.ca
RALI-DIRO
Université de Montréal
C.P. 6128, Succ Centre-Ville
Montréal (Québec) Canada
H3C 3J7

Atefeh Farzindar
farzindar@nlptechnologies.ca
NLP Technologies Inc.
52 Le Royer
Montréal
(Québec) Canada
H2Y 1W7

Abstract

While the automatic translation of tweets has already been investigated in different scenarios, we are not aware of any attempt to translate tweets created by government agencies. In this study, we report the experimental results we obtained when translating 12 Twitter feeds published by agencies and organizations of the government of Canada, using a state-of-the art Statistical Machine Translation (SMT) engine as a black box translation device. We mine parallel web pages linked from the URLs contained in English-French pairs of tweets in order to create tuning and training material. For a Twitter feed that would have been otherwise difficult to translate, we report significant gains in translation quality using this strategy. Furthermore, we give a detailed account of the problems we still face, such as hashtag translation as well as the generation of tweets of legal length.

1 Introduction

Twitter is currently one of the most popular online social networking service after Facebook, and is the fastest-growing, with the half-a-billion user mark reached in June 2012.[1] According to Twitter's blog, no less than 65 millions of tweets are published each day, mostly in a single language (40% in English). This hinders the spread of information, a situation witnessed for instance during the Arab Spring.

Solutions for disseminating tweets in different languages have been designed. One solution consists in manually translating tweets, which of course is only viable for a very specific subset of the material appearing on Twitter. For instance, the non-profit organization *Meedan*[2] has been founded in order to organize volunteers willing to translate tweets written in Arabic on Middle East issues. Another solution consists in using machine translation. Several portals are facilitating this,[3] mainly by using Google's machine translation API.

Curiously enough, few studies have focused on the automatic translation of text produced within social networks, even though a growing number of these studies concentrate on the automated processing of messages exchanged on social networks. See (Gimpel et al., 2011) for a recent review of some of them.

Some effort has been invested in translating short text messages (SMSs). Notably, Munro (2010) describes the service deployed by a consortium of volunteer organizations named "Mission 4636" during the earthquake that struck Haiti in January 2010. This service routed SMSs alerts reporting trapped people and other emergencies to a set of volunteers who translated Haitian Creole SMSs into English, so that primary emergency responders could understand them. In Lewis (2010), the authors describe how the Microsoft translation team developed a statistical translation engine (Haitian Creole into English) in as little as 5 days, during the same tragedy.

[1] http://semiocast.com/publications/2012_07_30_Twitter_reaches_half_a_billion_accounts_140m_in_the_US

[2] http://news.meedan.net/

[3] http://www.aboutonlinetips.com/twitter-translation-tools/

Proceedings of the Workshop on Language in Social Media (LASM 2013), pages 80–89,
Atlanta, Georgia, June 13 2013.

Jehl (2010) addresses the task of translating English tweets into German. She concludes that the proper treatment of unknown words is of the utmost importance and highlights the problem of producing translations of up to 140 characters, the upper limit on tweet lengths. In (Jehl et al., 2012), the authors describe their efforts to collect bilingual tweets from a stream of tweets acquired programmatically, and show the impact of such a collection on developing an Arabic-to-English translation system.

The present study participates in the effort for the dissemination of messages exchanged over Twitter in different languages, but with a very narrow focus, which we believe has not been addressed specifically yet: Translating tweets written by government institutions. What sets these messages apart is that, generally speaking, they are written in a proper language (without which their credibility would presumably be hurt), while still having to be extremely brief to abide by the ever-present limit of 140 characters. This contrasts with typical social media texts in which a large variability in quality is observed (Agichtein et al., 2008).

Tweets from government institutions can also differ somewhat from some other, more informal social media texts in their intended audience and objectives. Specifically, such tweet feeds often attempt to serve as a credible source of timely information presented in a way that engages members of the lay public. As such, translations should present a similar degree of credibility, ease of understanding, and ability to engage the audience as in the source tweet—all while conforming to the 140 character limits.

This study attempts to take these matters into account for the task of translating Twitter feeds emitted by Canadian governmental institutions. This could prove very useful, since more than 150 Canadian agencies have official feeds. Moreover, while only counting 34 million inhabitants, Canada ranks fifth in the number of Twitter users (3% of all users) after the US, the UK, Australia, and Brazil.[4] This certainly explains why Canadian governments, politicians and institutions are making an increasing use of this social network service. Given the need of Canadian governmental institutions to disseminate information in both official languages (French and English), we see a great potential value in targeted computer-aided translation tools, which could offer a significant reduction over the current time and effort required to manually translate tweets.

We show that a state-of-the-art SMT toolkit, used off-the-shelf, and trained on out-domain data is unsurprisingly not up to the task. We report in Section 2 our efforts in mining bilingual material from the Internet, which proves eventually useful in significantly improving the performance of the engine. We test the impact of simple adaptation scenarios in Section 3 and show the significant improvements in BLEU scores obtained thanks to the corpora we mined. In Section 4, we provide a detailed account of the problems that remain to be solved, including the translation of hashtags (#-words) omnipresent in tweets and the generation of translations of legal lengths. We conclude this work-in-progress and discuss further research avenues in Section 5.

2 Corpora

2.1 Bilingual Twitter Feeds

An exhaustive list of Twitter feeds published by Canadian government agencies and organizations can be found on the GOV.PoliTWiTTER.ca web site.[5] As of this writing, 152 tweet feeds are listed, most of which are available in both French and English, in keeping with the Official Languages Act of Canada. We manually selected 20 of these feed pairs, using various exploratory criteria, such as their respective government agency, the topics being addressed and, importantly, the perceived degree of parallelism between the corresponding French and English feeds.

All the tweets of these 20 feed pairs were gathered using Twitter's Streaming API on 26 March 2013. We filtered out the tweets that were marked by the API as retweets and replies, because they rarely have an official translation. Each pair of filtered feeds was then aligned at the tweet level in order to create bilingual tweet pairs. This step was facilitated by the fact that timestamps are assigned to each tweet. Since a tweet and its translation are typi-

[4] http://www.techvibes.com/blog/how-canada-stacks-up-against-the-world-on-twitter-2012-10-17

[5] http://gov.politwitter.ca/directory/network/twitter

Tweets	URLs	*mis.*	*probs*	*sents*
▷ *HealthCanada*				
1489	995	1	252	78,847
▷ *DFAIT_MAECI – Foreign Affairs and Int'l Trade*				
1433	65	0	1081	10,428
▷ *canadabusiness*				
1265	623	1	363	138,887
▷ *pmharper – Prime Minister Harper*				
752	114	2	364	12,883
▷ *TCS_SDC – Canadian Trade Commissioner Service*				
694	358	1	127	36,785
▷ *Canada_Trade*				
601	238	1	92	22,594
▷ *PHAC_GC – Public Health Canada*				
555	140	0	216	14,617
▷ *cida_ca – Canadian Int'l Development Agency*				
546	209	2	121	18,343
▷ *LibraryArchives*				
490	92	1	171	6,946
▷ *CanBorder – Canadian Border matters*				
333	88	0	40	9,329
▷ *Get_Prepared – Emergency preparedness*				
314	62	0	11	10,092
▷ *Safety_Canada*				
286	60	1	17	3,182

Table 1: Main characteristics of the Twitter and URL corpora for the 12 feed pairs we considered. The (English) feed name is underlined, and stands for the pair of feeds that are a translation of one another. When not obvious, a short description is provided. Each feed name can be found as is on Twitter. See Sections 2.1 and 2.3 for more.

cally issued at about the same time, we were able to align the tweets using a dynamic programming algorithm miminizing the total time drift between the English and the French feeds. Finally, we tokenized the tweets using an adapted version of `Twokenize` (O'Connor et al., 2010), accounting for the hashtags, usernames and urls contained in tweets.

We eventually had to narrow down further the number of feed pairs of interest to the 12 most prolific ones. For instance, the feed pair *PassportCan*[6] that we initially considered contained only 54 pairs of English-French tweets after filtering and alignment, and was discarded because too scarce.

[6] `https://twitter.com/PassportCan`

Did you know it's best to test for #radon in the fall/winter? `http://t.co/CDubjbpS` #health #safety
L'automne/l'hiver est le meilleur moment pour tester le taux de radon. `http://t.co/4NJWJmuN` #santé #sécurite

Figure 1: Example of a pair of tweets extracted from the feed pair *HealthCanada*.

The main characteristics of the 12 feed pairs we ultimately retained are reported in Table 1, for a total of 8758 tweet pairs. The largest feed, in terms of the number of tweet pairs used, is that of Health Canada[7] with over 1 489 pairs of retained tweets pairs at the time of acquisition. For reference, that is 62% of the 2 395 "raw" tweets available on the English feed, before filtering and alignment. An example of a retained pair of tweets is shown in Figure 1. In this example, both tweets contain a shortened url alias that (when expanded) leads to webpages that are parallel. Both tweets also contain so-called hashtags (#-words): 2 of those are correctly translated when going from English to French, but the hashtag *#radon* is not translated into a hashtag in French, instead appearing as the plain word *radon*, for unknown reasons.

2.2 Out-of-domain Corpora: Parliament Debates

We made use of two different large corpora in order to train our baseline SMT engines. We used the 2M sentence pairs of the Europarl version 7 corpus.[8] This is a priori an out-of-domain corpus, and we did not expect much of the SMT system trained on this dataset. Still, it is one of the most popular parallel corpus available to the community and serves as a reference.

We also made use of 2M pairs of sentences we extracted from an in-house version of the Canadian Hansard corpus. This material is not completely out-of-domain, since the matters addressed within the Canadian Parliament debates likely coincide to some degree with those tweeted by Canadian institutions. The main characteristics of these two corpora are reported in Table 2. It is noteworthy that while both

[7] `https://twitter.com/HealthCanada`

[8] `http://www.statmt.org/europarl/`

Corpus		sents	tokens	types	s length
hansard	en	2M	27.1M	62.2K	13.6
hansard	fr	2M	30.7M	82.2K	15.4
europarl	en	2M	55.9M	94.5K	28.0
europarl	fr	2M	61.6M	129.6K	30.8

Table 2: Number of sentence pairs, token and token types in the out-of-domain training corpora we used. s length stands for the average sentence length, counted in tokens.

corpora contain an equal number of sentence pairs, the average sentence length in the Europarl corpus is much higher, leading to a much larger set of tokens.

2.3 In-domain Corpus: URL Corpus

As illustrated in Figure 1, many tweets act as "teasers", and link to web pages containing (much) more information on the topic the tweet feed typically addresses. Therefore, a natural way of adapting a corpus-driven translation engine consists in mining the parallel text available at those urls.

In our case, we set aside the last 200 tweet pairs of each feed as a test corpus. The rest serves as the url-mining corpus. This is necessary to avoid testing our system on test tweets whose URLs have contributed to the training corpus.

Although simple in principle, this data collection operation consists in numerous steps, outlined below:

1. Split each feed pair in two: The last 200 tweet pairs are set aside for testing purposes, the rest serves as the url-mining corpus used in the following steps.

2. Isolate urls in a given tweet pair using our tokenizer, adapted to handle Twitter text (including urls).

3. Expand shortened urls. For instance, the url in the English example of Figure 1 would be expanded into `http://www.hc-sc.gc.ca/ewh-semt/radiation/radon/testing-analyse-eng.php`, using the expansion service located at the domain `t.co`. There are 330 such services on the Web.

4. Download the linked documents.

5. Extract all text from the web pages, without targeting any content in particular (the site menus, breadcrumb, and other elements are therefore retained).

6. Segment the text into sentences, and tokenize them into words.

7. Align sentences with our in-house aligner.

We implemented a number of restrictions during this process. We did not try to match urls in cases where the number of urls in each tweet differed (see column *mis.*—mismatches—in Table 1). The column *probs.* (problems) in Table 1 shows the count of url pairs whose content could not be extracted. This happened when we encountered urls that we could not expand, as well as those returning a 404 HTTP error code. We also rejected urls that were identical in both tweets, because they obviously could not be translations. We also filtered out documents that were not in html format, and we removed document pairs where at least one document was difficult to convert into text (e.g. because of empty content, or problematic character encoding). After inspection, we also decided to discard sentences that counted less than 10 words, because shorter sentences are too often irrelevant website elements (menu items, breadcrumbs, copyright notices, etc.).

This 4-hour long operation (including download) yielded a number of useful web documents and extracted sentence pairs reported in Table 1 (columns URLs and *sents* respectively). We observed that the density of url pairs present in pairs of tweets varies among feeds. Still, for all feeds, we were able to gather a set of (presumably) parallel sentence pairs.

The validity of our extraction process rests on the hypothesis that the documents mentioned in each pair of urls are parallel. In order to verify this, we manually evaluated (a posteriori) the parallelness of a random sample of 50 sentence pairs extracted for each feed. Quite fortunately, the extracted material was of excellent quality, with most samples containing all perfectly aligned sentences. Only *canadabusiness*, *LibraryArchives* and *CanBorder* counted a single mistranslated pair. Clearly, the websites of the Canadian institutions we mined are translated with great care and the tweets referring to them are meticulously translated in terms of content links.

3 Experiments

3.1 Methodology

All our translation experiments were conducted with Moses' EMS toolkit (Koehn et al., 2007), which in turn uses gizapp (Och and Ney, 2003) and SRILM (Stolcke, 2002).

As a test bed, we used the 200 bilingual tweets we acquired that were not used to follow urls, as described in Sections 2.1 and 2.3. We kept each feed separate in order to measure the performance of our system on each of them. Therefore we have 12 test sets.

We tested two configurations: one in which an out-of-domain translation system is applied (without adaptation) to the translation of the tweets of our test material, another one where we allowed the system to look at in-domain data, either at training or at tuning time. The in-domain material we used for adapting our systems is the URL corpus we described in section 2.3. More precisely, we prepared 12 tuning corpora, one for each feed, each containing 800 heldout sentence pairs. The same number of sentence pairs was considered for out-domain tuning sets, in order not to bias the results in favor of larger sets. For adaptation experiments conducted at training time, all the URL material extracted from a specific feed (except for the sentences of the tuning sets) was used. The language model used in our experiments was a 5-gram language model with Kneser-Ney smoothing.

It must be emphasized that there is no tweet material in our training or tuning sets. One reason for this is that we did not have enough tweets to populate our training corpus. Also, this corresponds to a realistic scenario where we want to translate a Twitter feed without first collecting tweets from this feed.

We use the BLEU metric (Papineni et al., 2002) as well as word-error rate (WER) to measure translation quality. A good translation system maximizes BLEU and minimizes WER. Due to initially poor results, we had to refine the tokenizer mentioned in Section 2.1 in order to replace urls with serialized placeholders, since those numerous entities typically require rule-based translations. The BLEU and WER scores we report henceforth were computed on such lowercased, tokenized and serialized texts, and did not incur penalties that would have otherwise been caused by the non-translation of urls (unknown tokens), for instance.

train	tune	*canadabusiness*		*DFAIT_MAECI*	
fr→en		wer	bleu	wer	bleu
hans	hans	59.58	21.16	61.79	19.55
hans	in	58.70	21.35	60.73	20.14
euro	euro	64.24	15.88	62.90	17.80
euro	in	63.23	17.48	60.58	21.23
en→fr		wer	bleu	wer	bleu
hans	hans	62.42	21.71	64.61	21.43
hans	in	61.97	22.92	62.69	22.00
euro	euro	64.66	19.52	63.91	21.65
euro	in	64.61	18.84	63.56	22.31

Table 3: Performance of generic systems versus systems adapted at tuning time for two particular feeds. The tune corpus "in" stands for the URL corpus specific to the feed being translated. The tune corpora "hans" and "euro" are considered out-of-domain for the purpose of this experiment.

3.2 Translation Results

Table 3 reports the results observed for the two main configurations we tested, in both translation directions. We show results only for two feeds here: *canadabusiness*, for which we collected the largest number of sentence pairs in the URL corpus, and *DFAIT_MAECI* for which we collected very little material. For *canadabusiness*, the performance of the system trained on Hansard data is higher than that of the system trained on Europarl (Δ ranging from 2.19 to 5.28 points of BLEU depending on the configuration considered). For *DFAIT_MAECI*, suprisingly, Europarl gives a better result, but by a more narrow margin (Δ ranging from 0.19 to 1.75 points of BLEU). Both tweet feeds are translated with comparable performance by SMT, both in terms of BLEU and WER. When comparing BLEU performances based solely on the tuning corpus used, the in-domain tuning corpus created by mining urls yields better results than the out-domain tuning corpus seven times out of eight for the results shown in Table 3.

The complete results are shown in Figure 2, showing BLEU scores obtained for the 12 feeds we considered, when translating from English to French. Here, the impact of using in-domain data to tune

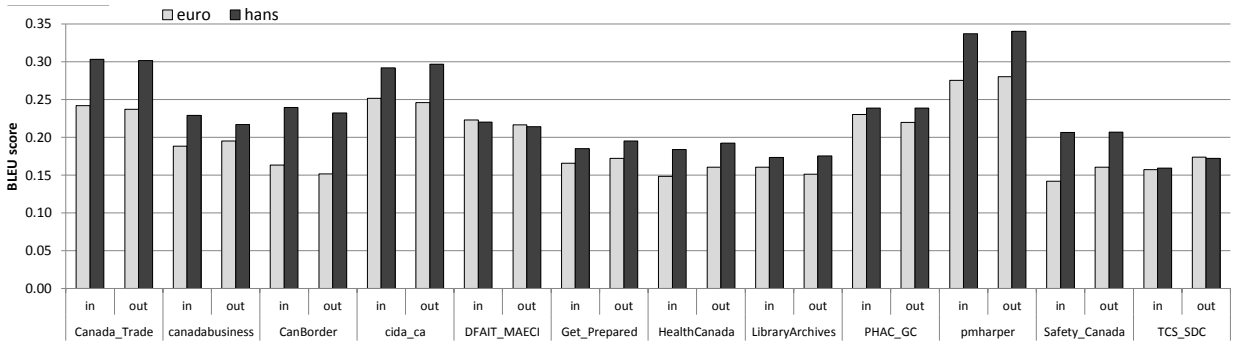

Figure 2: BLEU scores measured on the 12 feed pairs we considered for the English-to-French translation direction. For each tweet test corpus, there are 4 results: a dark histogram bar refers to the Hansard training corpus, while a lighter grey bar refers to an experiment where the training corpus was Europarl. The "in" category on the x-axis designates an experiment where the tuning corpus was in-domain (URL corpus), while the "out" category refers to an out-of-domain tuning set. The out-of-domain tuning corpus is Europarl or Hansard, and always matches the nature of training corpora.

the system is hardly discernible, which in a sense is good news, since tuning a system for each feed is not practical. The Hansard corpus almost always gives better results, in keeping with its status as a corpus that is not so out-of-domain as Europarl, as mentioned above. The results for the reverse translation direction show the same trends.

In order to try a different strategy than using only tuning corpora to adapt the system, we also investigated the impact of training the system on a mix of out-of-domain and in-domain data. We ran one of the simplest adaptation scenarios where we concatenated the in-domain material (train part of the URL corpus) to the out-domain one (Hansard corpus) for the two feeds we considered in Table 3. The results are reported in Table 4.

We measured significant gains both in WER and BLEU scores in conducting training time versus tuning time adaptation, for the *canadabusiness* feed (the largest URL corpus). For this corpus, we observe an interesting gain of more than 6 absolute points in BLEU scores. However, for the *DFAIT_MAECI* (the smallest URL corpus) we note a very modest loss in translation quality when translating from French and a significant gain in the other translation direction. These figures could show that mining parallel sentences present in URLs is a fruitful strategy for adapting the translation engine for feeds like *canadabusiness* that display poor performance otherwise, without harming the translation quality for feeds that perform reasonably well without additional resources. Unfortunately, it suggests that retraining a system is required for better performance, which might hinder the deployment of a standalone translation engine. Further research needs to be carried out to determine how many tweet pairs must be used in a parallel URL corpus in order to get a sufficiently good in-domain corpus.

Train corpus	WER	BLEU
	fr→en	
hans+*canbusiness*	53.46 *(-5.24)*	27.60 *(+6.25)*
hans+*DFAIT*	60.81 *(+0.23)*	20.83 *(-0.40)*
	en→fr	
hans+*canbusiness*	57.07 *(-4.90)*	26.26 *(+3.34)*
hans+*DFAIT*	61.80 *(-0.89)*	24.93 *(+2.62)*

Table 4: Performance of systems trained on a concatenation of out-of-domain and in-domain data. All systems were tuned on in-domain data. Absolute gains are shown in parentheses, over the best performance achieved so far (see Table 3).

4 Analysis

4.1 Translation output

Examples of translations produced by the best system we trained are reported in Figure 3. The first translation shows a case of an unknown French word (*soumissionnez*). The second example illustrates

a typical example where the hashtags should have been translated but were left unchanged. The third example shows a correct translation, except that the length of the translation (once the text is detokenized) is over the size limit allowed for a tweet. Those problems are further analyzed in the remaining subsections.

4.2 Unknown words

Unknown words negatively impact the quality of MT output in several ways. First, they typically appear untranslated in the system's output (we deemed most appropriate this last resort strategy). Secondly, they perturb the language model, which often causes other problems (such as dubious word ordering). Table 5 reports the main characteristics of the words from all the tweets we collected that were not present in the Hansard train corpus.

The out-of-vocabulary rate with respect to token types hovers around 33% for both languages. No less than 42% (resp. 37%) of the unknown English (resp. French) token types are actually hashtags. We defer their analysis to the next section. Also, 15% (resp. 10%) of unknown English token types are user names (@user), which do not require translation.

	English	French
tweet tokens	153 234	173 921
tweet types	13 921	15 714
OOV types	4 875 *(35.0%)*	5 116 *(32.6%)*
▷ hashtag types	2 049 *(42.0%)*	1 909 *(37.3%)*
▷ @user types	756 *(15.5%)*	521 *(10.2%)*

Table 5: Statistics on out-of-vocabulary token types.

We manually analyzed 100 unknown token types that were not hashtags or usernames and that did not contain any digit. We classified them into a number of broad classes whose distributions are reported in Table 6 for the French unknown types. A similar distribution was observed for English unknown types. While we could not decide of the nature of 21 types without their context of use (line *?type*), we frequently observed English types, as well as acronyms and proper names. A few unknown types result from typos, while many are indeed true French types unseen at training time (row labeled *french*), some of which being very specific (*term*). Amusingly, the French verbal neologism *twitter* (*to tweet*) is unknown to the Hansard corpus we used.

french	26	*sautez*, *perforateurs*, *twitter*
english	22	*successful*, *beauty*
?types	21	*bumbo*, *tra*
name	11	*absorbica*, *konzonguizi*
acronym	7	*hna*, *rnc*
typo	6	*gazouilli*, *pendan*
term	3	*apostasie*, *sibutramine*
foreign	2	*aanischaaukamikw*, *aliskiren*
others	2	*francophones*URL

Table 6: Distribution of 100 unknown French token types (excluding hashtags and usernames).

4.3 Dealing with Hashtags

We have already seen that translating the text in hashtags is often suitable, but not always. Typically, hashtags in the middle of a sentence are to be translated, while those at the end typically should not be. A model should be designed for learning when to translate an hashtag or not. Also, some hashtags are part of the sentence, while others are just (semantic) tags. While a simple strategy for translating hashtags consists in removing the # sign at translation time, then restoring it afterwards, this strategy would fail in a number of cases that require segmenting the text of the hashtag first. Table 7 reports the percentage of hashtags that should be segmented before being translated, according to a manual analysis we conducted over 1000 hashtags in both languages we considered. While many hashtags are single words, roughly 20% of them are not and require segmentation.

4.4 Translating under size constraints

The 140 character limit Twitter imposes on tweets is well known and demands a certain degree of concision even human users find sometimes bothersome. For machine output, this limit becomes a challenging problem. While there exists plain–but inelegant–workarounds[9], there may be a way to *produce* tweet translations that are themselves Twitter-ready. (Jehl,

[9]The service eztweets.com splits long tweets into smaller ones; twitlonger.com tweets the beginning of a long message,

SRC:	vous soumissionnez pour obtenir de gros contrats ? voici 5 pratiques exemplaires à suivre . URL
TRA:	you soumissionnez big contracts for best practices ? here is 5 URL to follow .
REF:	bidding on big contracts ? here are 5 best practices to follow . URL
SRC:	avis de #santépublique : maladies associées aux #salmonelles et à la nourriture pour animaux de compagnie URL #rappel
TRA:	notice of #santépublique : disease associated with the #salmonelles and pet food #rappel URL
REF:	#publichealth notice : illnesses related to #salmonella and #petfood URL #recall
SRC:	des haïtiens de tous les âges , milieux et métiers témoignent de l' aide qu' ils ont reçue depuis le séisme . URL #haïti
TRA:	the haitian people of all ages and backgrounds and trades testify to the assistance that they have received from the earthquake #haïti URL .
REF:	#canada in #haiti : haitians of all ages , backgrounds , and occupations tell of the help they received . URL

Figure 3: Examples of translations produced by an engine trained on a mix of in- and out-of-domain data.

w.	en	fr	example
1	76.5	79.9	*intelligence*
2	18.3	11.9	*gender equality*
3	4.0	6.0	*africa trade mission*
4	1.0	1.4	*closer than you think*
5	0.2	0.6	*i am making a difference*
6	–	0.2	*fonds aide victime sécheresse afrique est*

Table 7: Percentage of hashtags that require segmentation prior to translation. *w.* stands for the number of words into which the hashtag text should be segmented.

2010) pointed out this problem and reported that 3.4% of tweets produced were overlong, when translating from German to English. The reverse directions produced 17.2% of overlong German tweets. To remedy this, she tried modifying the way BLEU is computed to penalize long translation during the tuning process, with BLEU scores worse than simply truncating the illegal tweets. The second strategy the author tried consisted in generating n-best lists and mining them to find legal tweets, with encouraging results (for $n = 30\,000$), since the number of overlong tweets was significantly reduced while leaving BLEU scores unharmed.

In order to assess the importance of the problem for our system, we measured the lengths of tweets that a system trained like hans+*canbusiness* in Table 4 (a mix of in- and out-of-domain data) could produce. This time however, we used a larger test set counting 498 tweets. To measure the lengths of their translations, we first had to detokenize the translations produced, since the limitation applies to "natural" text only. For each URL serialized token, we counted 18 characters, the average length of a (shortened) url in a tweet. When translating from French to English, the 498 translations had lengths ranging from 45 to 138 characters; hence, they were all legal tweets. From English to French, however, the translations are longer, and range from 32 characters to 223 characters, with 22.5% of them overlong.

One must recall that in our experiments, no tweets were seen at training or tuning time, which explains why the rate of translations that do not meet the limit is high. This problem deserves a specific treatment for a system to be deployed. One interesting solution already described by (Jehl, 2010) is to mine the n-best list produced by the decoder in order to find the first candidate that constitutes a legal tweet. This candidate is then picked as the translation. We performed this analysis on the *canadabusiness* output described earlier, from English to French. We used n =1, 5, 10, 20, 50, 100, 200, 500, 1000, 5000, 10000, 30000 and computed the resulting BLEU scores and remaining percentage of overlong tweets. The results are shown in Figure 4. The results clearly show that the n-best list does contain alternate candidates when the best one is too long. Indeed, not only do we observe that the percentage of remaining illegal tweets can fall steadily (from 22.4% to 6.6% for $n = 30\,000$) as we dig deeper into the list, but also the BLEU score stays unharmed, showing even a slight improvement, from an ini-

and provides a link to read the remainder. One could also simply truncate an illegal tweet and hope for the best...

tial 26.16 to 26.31 for $n = 30\,000$. This counter-intuitive result in terms of BLEU is also reported in (Jehl, 2010) and is probably due to a less harsh brevity penalty by BLEU on shorter candidates.

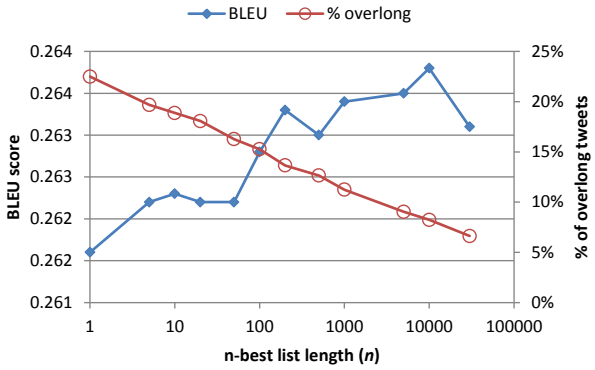

Figure 4: BLEU scores and percentage of overlong tweets when mining the n-best list for legal tweets, when the first candidate is overlong. The BLEU scores (diamond series) should be read off the left-hand vertical axis, while the remaining percentage of illegal tweets (circle series) should be read off the right-hand axis.

5 Discussion

We presented a number of experiments where we translated tweets produced by Canadian governments institutions and organizations. Those tweets have the distinguishing characteristic (in the Twittersphere) of being written in proper English or French. We show that mining the urls mentioned in those tweets for parallel sentences can be a fruitful strategy for adapting an out-of-domain translation engine to this task, although further research could show other ways of using this resource, whose quality seems to be high according to our manual evaluation. We also analyzed the main problems that remain to be addressed before deploying a useful system.

While we focused here on acquiring useful corpora for adapting a translation engine, we admit that the adaptation scenario we considered is very simplistic, although efficient. We are currently investigating the merit of different methods to adaptation (Zhao et al., 2004; Foster et al., 2010; Daume III and Jagarlamudi, 2011; Razmara et al., 2012; Sankaran et al., 2012).

Unknown words are of concern, and should be dealt with appropriately. The serialization of urls was natural, but it could be extended to usernames. The latter do not need to be translated, but reducing the vocabulary is always desirable when working with a statistical machine translation engine. One interesting subcategories of out-of-vocabulary tokens are hashtags. According to our analysis, they require segmentation into words before being translated in 20% of the cases. Even if they are transformed into regular words (#radon→radon or #genderequality→gender equality), however, it is not clear at this point how to detect if they are used like normally-occurring words in a sentence, as in (#radon is harmful) or if they are simply tags added to the tweet to categorize it.

We also showed that translating under size constraints can be handled easily by mining the n-best list produced by the decoder, but only up to a point. A remaining 6% of the tweets we analyzed in detail could not find a shorter version. Numerous ideas are possible to alleviate the problem. One could for instance modify the logic of the decoder to penalize hypotheses that promise to yield overlong translations. Another idea would be to manually inspect the strategies used by governmental agencies on Twitter when attempting to shorten their messages, and to select those that seem acceptable and implementable, like the suppression of articles or the use of authorized abbreviations.

Adapting a translation pipeline to the very specific world of governmental tweets therefore poses multiple challenges, each of which can be addressed in numerous ways. We have reported here the results of a modest but fertile subset of these adaptation strategies.

Acknowledgments

This work was funded by a grant from the Natural Sciences and Engineering Research Council of Canada. We also wish to thank Houssem Eddine Dridi for his help with the Twitter API.

References

Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. 2008. Finding high-quality content in social media. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, WSDM '08, pages 183–194.

Hal Daume III and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *49th ACL*, pages 407–412, Portland, Oregon, USA, June.

George Foster, Cyril Goutte, and Roland Kuhn. 2010. Discriminative instance weighting for domain adaptation in statistical machine translation. In *EMNLP*, pages 451–459, Cambridge, MA, October.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *ACL (Short Papers)*, pages 42–47.

Laura Jehl, Felix Hieber, and Stefan Riezler. 2012. Twitter translation using translation-based cross-lingual retrieval. In *7th Workshop on Statistical Machine Translation*, pages 410–421, Montréal, June.

Laura Jehl. 2010. Machine translation for twitter. Master's thesis, School of Philosophie, Psychology and Language Studies, University of Edinburgh.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-burch, Richard Zens, Rwth Aachen, Alexandra Constantin, Marcello Federico, Nicola Bertoldi, Chris Dyer, Brooke Cowan, Wade Shen, Christine Moran, and Ondřej Bojar. 2007. Moses: Open source toolkit for statistical machine translation. pages 177–180.

William D. Lewis. 2010. Haitian creole: How to build and ship an mt engine from scratch in 4 days, 17 hours, & 30 minutes. In *EAMT*, Saint-Raphael.

Robert Munro. 2010. Crowdsourced translation for emergency response in Haiti: the global collaboration of local knowledge. In *AMTA Workshop on Collaborative Crowdsourcing for Translation*, Denver.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Comput. Linguist.*, 29(1):19–51, March.

Brendan O'Connor, Michel Krieger, and David Ahn. 2010. TweetMotif: Exploratory Search and Topic Summarization for Twitter. In William W. Cohen, Samuel Gosling, William W. Cohen, and Samuel Gosling, editors, *ICWSM*. The AAAI Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei J. Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Majid Razmara, George Foster, Baskaran Sankaran, and Anoop Sarkar. 2012. Mixing multiple translation models in statistical machine translation. In *Proceedings of the 50th ACL*, Jeju, Republic of Korea, jul.

Baskaran Sankaran, Majid Razmara, Atefeh Farzindar, Wael Khreich, Fred Popowich, and Anoop Sarkar. 2012. Domain adaptation techniques for machine translation and their evaluation in a real-world setting. In *Proceedings of 25th Canadian Conference on Artificial Intelligence*, Toronto, Canada, may.

Andreas Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proceedings of ICSLP*, volume 2, pages 901–904, Denver, USA.

Bing Zhao, Matthias Eck, and Stephan Vogel. 2004. Language model adaptation for statistical machine translation with structured query models. In *20th COLING*.