Hashtag Occurrences, Layout and Translation: A Corpus-driven Analysis of Tweets Published by the Canadian Government

Fabrizio Gotti[1], Philippe Langlais[1], Atefeh Farzindar[2]

[1] RALI, Université de Montréal, Pav. André-Aisenstadt, C.P. 6128, Succ. Centre-Ville, Montréal, Canada, H3C 3J7

[2] NLP Technologies Inc., 52 Le Royer, Montréal, Canada, H2Y 1W7

gottif@iro.umontreal.ca, felipe@iro.umontreal.ca, farzindar@nlptechnologies.ca

Abstract

In this article, we present an aligned bilingual corpus of 8758 tweet pairs in French and English, derived from 12 Canadian government agencies. Hashtags account for 6% to 8% of all tokens, and exhibit a Zipfian distribution. They appear in either a tweet's prologue, announcing its topic, or in the tweet's text in lieu of traditional words, or in an epilogue. Hashtags are words prefixed with a pound sign in 80% of the cases. The rest is mostly multiword hashtags, for which we describe a simple segmentation algorithm. A manual analysis of the bilingual alignment of 5000 hashtags shows that 5% (French) to 18% (English) of them don't have a counterpart in their containing tweet's translation. This analysis further shows that 80% of multiword hashtags are correctly translated by humans, and that the mistranslation of the rest may be due to incomplete translation directives regarding social media. We show how these resources and their analysis can guide the design of a statistical machine translation pipeline, and its evaluation. A baseline system implementing a tweet-specific tokenizer yields promising results. The system is improved by translating epilogues, prologues, and text separately. We attempt to feed the SMT engine with the original hashtag and some alternatives ("dehashed" version or a segmented version of multiword hashtags), but translation quality improves at the cost of hashtag recall.

Keywords: Twitter hashtags, bilingual corpus, machine translation

1. Introduction

The meteoric rise of Twitter to reach the place of second most popular social networking site in the world has drawn the attention of the natural language processing community, focusing on topics such as sentiment detection (Roberts et al., 2012), opinion mining and machine translation (Jehl, 2010; Jehl et al., 2012).

One of the prominent features of tweets is hashtags. Hashtags are words or phrases consisting of alphanumeric characters prefixed with the pound sign (#), e.g. `#health` or `#49MillionBeliebers`. Authors use hashtags liberally within tweets to mark them as belonging to a particular topic, and hashtags can serve to group messages belonging to a topic. Twitter.com features a search engine that can show in real-time the activity pertaining to a hashtag, i.e. the tweets, news, images, videos, etc. associated with such a topic in the so-called "Twittersphere". Hashtags further provide a way to label and monitor emerging trends, be they local or worldwide.

They are indeed a very interesting form of metadata and are featured in other microblogging and social networking services, such as Facebook and the very popular Chinese platform Sina Weibo. Previous studies have examined how hashtags can be automatically suggested, mined, or translated. The latter task is complicated by the facts that hashtags occur in various positions within a post, are often named entities, and may be agglutinations of several words or non-alphabetic characters.

To further the study of this phenomenon, we present here a bilingual Twitter corpus extracted from tweets issued by the government of Canada. To our knowledge, this is one of the first such bilingual resources made available to the community. Our goal in creating this resource is to provide statistical knowledge about hashtags "in the wild" and to show how this resource and its analysis can guide design choices in creating a statistical machine translation (SMT) engine adapted to the translation of tweets from English to French and vice-versa. It is our hope that the resources presented here could help others in understanding hashtag occurrences and nature in tweets, and in perfecting and evaluating a machine translation pipeline for the text containing them.

One important previous resource stems from a large-scale data-mining approach (Ling et al., 2013) performed on 1.6 G tweets and 65 M microblogging messages from Sina Weibo, which aimed at identifying single posts containing text in more than one language. Parallel text extracted from these messages allowed the authors to improve the translation quality of machine translation systems targeting the language pairs English-Chinese and English-Arabic[1]. Another study in the same vein (Jehl et al., 2012) showed that translation-based cross-language information retrieval can retrieve microblog messages across languages. They proved similar enough to be used to adapt a standard phrase-based SMT pipeline to the microblog domain.

We start by presenting the bilingual corpus we used in Section 2. In Section 3, we explore the frequency of occurrence of hashtags, their layout in tweets as well as their relationships to the "ordinary" vocabulary. In Section 4, we proceed to show how the humans who translated the tweets have managed to transfer hashtags from one language to another, and we show the problems they face in

[1] www.cs.cmu.edu/~lingwang/microtopia/

so doing. In Section 5, we show how the resources we provide can be used to improve the statistical machine translation of tweets and their hashtags. We conclude by highlighting the difficulties associated with hashtag translation, and more generally, with their handling by natural language processing applications.

2. A corpus of bilingual Twitter feeds

In keeping with the Official Languages Act of Canada, most official publications made by the Canadian government must be issued simultaneously in both English and French. This includes the material published on Twitter by more than 100 government agencies and bodies[2] and by some politicians, including the Prime Minister.

According to the result of our enquiries to a few of these agencies, tweet translation is handled by certified translators hired by the government, and is typically conducted from English to French. A qualitative analysis of the original tweets and their translation shows them to be of very high quality. Typically, we observed that most of these institutions have actually set up two Twitter accounts, one for each language, contrarily to some users who prefer to alternate French and English tweets on the same account[3], or to write single posts in two languages.

We downloaded 12 feed pairs using Twitter's Streaming API on 26 March 2013. We filtered out retweets and replies and aligned them at the tweet level to create bilingual tweet pairs. We carried out the alignment automatically by using the timestamps associated with each tweets. Indeed, a tweet and its translation are typically issued roughly at the same time. Therefore, it was possible to devise a dynamic programming algorithm whose cost function is proportional to the total time drift calculated between two feeds. We describe the corpus creation steps in more detail in (Gotti et al., 2013).

This bitext, which we call here **twitter-all**, counts 8758 tweet pairs. We describe it in detail in the next section. It is subdivided as follows:

- **twitter-test**: the last (most recent) 200 pairs of tweets from all 12 feeds constitute the test corpus, for a total of 2400 tweet pairs.
- **twitter-tune**: 758 pairs of tweets randomly selected from the rest
- **twitter-train**: the rest, counting 5600 tweet pairs.

An example of a tweet pair originating from Health Canada/Santé Canada is shown in Figure 1.

It is noteworthy that these resources are all tokenized and lowercased in this study. The encoding is always UTF-8. We tokenized and parsed the tweets using a slightly modified version of `Twokenize` (O'Connor et al., 2010), partly in order to better process French text and hashtags. This preprocessing allows us to quickly tokenize text and extract hashtags and URLs. We replaced all URLs with an arbitrary token shown as `<url>` henceforth.

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

[3] See for instance https://twitter.com/JustinTrudeau.

did you know it' s best to test for #radon in the fall / winter ? <url> #health #safety
l' automne / l' hiver est le meilleur moment pour tester le taux de radon. <url> #santé #sécurité

Figure 1: Example of a pair of tweets extracted from the bilingual feed pair Health Canada/Santé Canada, after tokenization

3. Occurrences, layout and nature of hashtags

In the following subsections, we show statistics for the corpus **twitter-all** (see Section 2).

3.1 Frequency of occurrence

A cursory glance at tweets indicates that hashtags are ubiquitous in Twitter feeds. However, if one is to design a natural language processing module aimed at handling them, having an idea of their frequency is useful to determine the effort that is reasonable to invest. We show the results in Table 1; the distribution of the number of hashtags is shown in Figure 2.

	English	French
# tweets	8758	8758
# tokens ("words" + hashtags)	142136	155153
# hashtags	11481	10254
# hashtag types	1922	1781
avg hashtags/tweet	1.31	1.17
% hashtags w.r.t. tokens	8.1%	6.6%
# tweets with at least one hashtag	5460	5137

Table 1: Statistics on hashtag use in corpus **twitter-all**

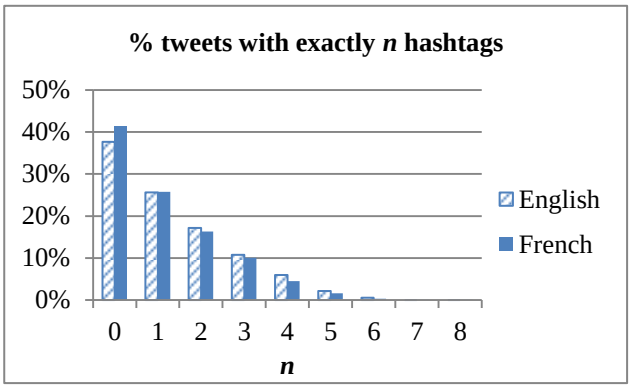

Figure 2: Distribution of the frequency of hashtag use in French and English tweets

Expectedly, the use of hashtags is quite frequent and deserves to be addressed, since it appears that, on average, more than 50% of tweets contain at least one hashtag. Moreover, no less than 8.1% of English tweet tokens consist of hashtags. This figure drops to 6.6% for the French corpus, due both to the relative wordiness of French compared to English and to the smaller number of

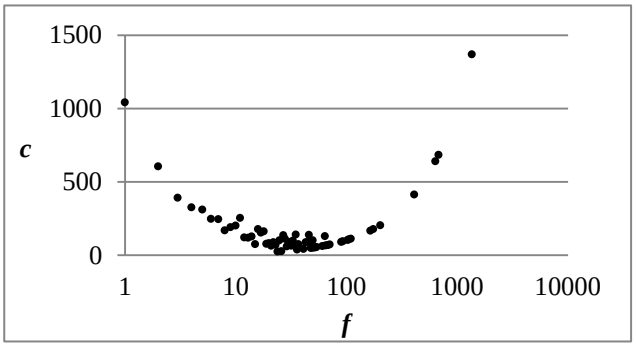

Figure 3: Count c of hashtags as a function of their frequency f for the English part of the **twitter-all** corpus. The x scale is logarithmic.

Figure 4: An original tweet and three possible regions for hashtags

hashtags in French. The latter observation immediately raises some questions about the faithfulness of the translation of hashtags from English to French. This is discussed further in Section 4.

From the data in Table 1, one could derive that the average frequency of occurrence for a given hashtag is 11481/1922 ≈ 5.97 occurrences in English, but this is misleading. The distribution is far from being uniform. We measured on the complete Twitter corpus (**twitter-all**) the frequency of occurrence of all hashtags found. We computed the following statistics: for each frequency f of hashtag found in the corpus, we calculated

$c = f \times$ number hashtags types having frequency f.

The results are shown in Figure 3, for English. The French curve is remarkably similar. Although not shown in the figure, hashtags exhibit a Zipfian distribution.

This indicates that numerous hashtag types are used very rarely and that a few hashtags types are used extensively across all tweets in the test corpus. For instance, there are 1041 English hashtags appearing only once (happaxes) and a single hashtag appearing 1367 times (`#canada`).

3.2 Layout

The location of hashtags in a tweet may hint at a way of handling the source text to spot them and process them in a principled manner. For instance, if they are isolated at the end of a tweet and serve no syntactical purpose in the sentence tweeted, then it may be best to translate them using a dictionary rather than a full-fledged machine translation engine.

Hashtags in our corpus seem to be distributed in three distinct regions, illustrated in Figure 4.

We recognize three (possibly empty) parts for a tweet:

- a **prologue**, announcing what the tweet is about
- the text itself, containing what we will call **inline hashtags**. These are regular words prefixed with a pound sign. In multi-word expressions, the individual words are usually concatenated.
- an **epilogue** (sometimes called a postscript), often containing URLs and tags added to further categorize the topic of the tweet.

This classification echoes in part the observations made in (Gimpel et al., 2011), where the Twitter part-of-speech tagger described distinguishes between "categorizing hashtags" occurring near the end of a post, and hashtags standing for other words (proper and common nouns, verbs, etc.) occurring within the text of the tweet.

Once again, we studied the frequency of the phenomenon by writing a Tweet splitter able to recognize heuristically the different parts of a Tweet, and we counted the number of hashtags contained in each part. The results are shown in Table 2.

The fact that about 40% of hashtags occur outside the tweet's text reinforces our intuition that those hashtags should be targeted specifically when processing them. The simplest idea is to adapt the sentence-splitting algorithm to isolate those epilogues and prologues from the rest of the text.

Moreover, since 87% of tweets contain an epilogue, potentially containing a URL, it may also be a good idea to treat this part with a specific algorithm. The URLs contained in these epilogues could be the first target of such a specific processing step.

	English	French
Number of tweets	8758	8758
% of tweets with a prologue	10.7%	10.1%
% of tweets with an epilogue	87.3%	86.5%
% of tweets with prologue & epilogue	10.4%	9.8%
Number of hashtags	11481	10254
% of hashtags in prologues	8.2%	8.7%
% of hashtags in epilogues	30.9%	28.9%
total % of hashtags in prologues and epilogues	39.1%	37.5%

Table 2: Distribution of hashtags in epilogues and prologues

3.3 Hashtags and OOVs

Hashtags complicate any automated language-processing step by occurring in odd places and by upsetting the natural order of words (section 3.2). They also can be unknown (out-of-vocabulary, OOV) in a language. We investigate this here.

Hashtags are by their very nature all unknown to the standard French or English vocabulary, since there are no words in these vocabularies that start with a pound sign. If finding a natural language equivalent to a given hashtag is only a matter of removing the pound sign, then we have found a way of (at the very least) normalizing this input for a machine. We cross-referenced the text of the hashtags (i.e. the hashtag without the pound sign) against French and English vocabularies as found in 3 M sentences of the Canadian Hansard corpus. The complete results are shown in Table 3.

	English	French
Number of hashtags	11481	10254
Number of hashtag types	1922	1781
% OOV hashtags stripped of the # sign	23.2%	20.6%
% OOV hashtag types stripped of #	16.7%	17.4%

Table 3: Percentage of unknown hashtags to English and French vocabularies of the Hansard corpus

The results clearly show that about 80% of the hashtags are actually in-vocabulary, since stripping them of their pound sign produces a word found in English or French. Although we do not have specific figures about the distribution of the remaining OOV hashtags, it clearly appears that the majority of them are multiword hashtags (for instance `#RaiseAReaderDay` or `#NouveauBrunswick`). We created a simple hashtag segmenting procedure backed by the corresponding language's vocabulary in order to approximate the proportion of the OOV hashtags that can be split into English (or French) words. The algorithm simply attempts to find out if an unknown hashtag can be split in substrings that are all known to the underlying vocabulary (including numbers). The segmentation procedure is language-dependent, because it relies on a given language's vocabulary, but is otherwise unchanged from language to language.

Some splits are evidently wrong (oversegmentation `[hoc, key, day]` or undersegmentation: `[recherchées, parla, sfc]`, where `parla` means "`bythe`") but we are merely interested in estimating whether OOV hashtags can be split into known words, to reduce the OOV rate. One improvement to this rather simple algorithm would be to take into account the case of the hashtags, leveraging the natural tendencies of some bloggers to use medial capitals (camel case) to mark word boundaries, as in `#TravelTuesday`.

	English	French
Number of hashtag types	1922	1781
% OOV hashtag types (from Table 2)	16.7%	17.4%
% OOV hashtag types after segmentation	3.7%	4.7%

Table 4: Percentage of unknown hashtags to "standard" English and French vocabularies, after automatic segmentation of multiword hashtags into simple words

The OOV percentage drops from 17% on average to about 4% on average (Table 4), which is quite encouraging, and would argue in favor of implementing multi-word segmentation for these hashtags.

4. Human translation of hashtags

In Section 3.1, we showed that there are hints of discrepancies between source- and target-language tweets regarding hashtags.

For the corpus **twitter-test** (2400 tweets – see Section 2), we manually aligned all hashtags contained in the tweets in order to determine how they are translated by humans. To help speed up this time-consuming task, we first built a simple bilingual dictionary of hashtags based on the corpus **twitter-all**.

Figure 1 shows an example of an unaligned hashtag (`#radon`) in English, with no hashtag counterpart in the translation of the tweet that contains it. This problem is also observed in the other translation direction.

Table 5 shows the statistics for aligned and unaligned hashtags. It clearly shows that misalignment in the reference is significant: out of 1376 tweet pairs with hashtags, 28.3% present a misalignment, i.e. at least one hashtag is lost in translation, in one direction or the other.

Statistic	**twitter-test**	
# tweet pairs	2400	
# tweet pairs with hashtags	1376	
# tweet pairs with > 0 unaligned hashtag	390 (28.3% of tweet pairs with #tags)	
	English corpus	**French corpus**
Nb hashtags	2682	2334
Nb hashtags unaligned	125 (4.7% of total)	473 (20.2% of total)

Table 5: Distribution of aligned and unaligned hashtags for the **twitter-test** corpus

This problem shows that even professional translators have hesitations and are inconsistent when translating hashtags, and the resource should be used carefully when automatically evaluating translation quality using such evaluation metric as BLEU (Papineni et al., 2002).

4.1 Hashtag translation and disposition

We identified three sections in a tweet where hashtags can appear: prologue, inline text and epilogue. We report here the alignment of the translation of hashtags appearing in each region. Moreover, we study hashtags that are aligned to hashtags not belonging to the same region (in the target tweet), like illustrated in Figure 5.

#media advisory regarding the beyond the border action plan <url>
#média : avis aux médias concernant le plan d' action frontalier <url>

Figure 5: Example of a hashtag (#media) belonging to the prologue in French, and inline in the English version

For each language (en or fr), for each section (pro, inline, epi), we show in Table 6 the percentage of hashtags found in **twitter-test** that are:

- **align-in:** the hashtag is aligned to a target-language hashtag in the same region
- **align-out:** the hashtag is aligned to a target-language hashtag in another region (for instance, a French hashtag inline aligned to an English counterpart in the prologue).
- **nil:** the hashtag is not translated (e.g. #radon in Figure 1).

	Region	Nb. of hashtags	align-in	align-out	nil
en	pro	221	88.2%	5.9%	5.9%
	inline	2433	80.5%	0.8%	18.7%
	epi	28	67.9%	17.9%	14.3%
	all-regions	2682	81.0%	1.4%	17.6%
fr	pro	202	96.5%	3.0%	0.5%
	inline	2099	93.3%	0.9%	5.9%
	epi	33	57.6%	39.4%	3.0%
	all-regions	2334	93.1%	1.6%	5.4%

Table 6: Distribution of hashtags aligned with their region (align-in), across regions (align-out) and not aligned at all (nil), for English and French hashtags, for each tweet region, in corpus **twitter-test**.

4.2 Translation of multiword hashtags

For multiword hashtags, there is inconsistency in the way humans translate governmental hashtags in **twitter-test**. We manually inspected the translations of the 373 occurrences of hashtags pairs and observed four different phenomena in multiword hashtag translations:

- **good** translation (80% of cases): faithful translation, for instance #stanleycup and #coupestanley
- **nil** (7 % of cases): the source hashtag has no hashtag counterpart in the translation (for instance #toysafety becomes la sécurité […] des jouets)
- **as-is** (8 % of cases): the hashtag is not translated, merely reproduced in the target language: #japanquake in both French and English. We found that the hashtags are all in English in this category.
- **part** (5 % of cases): one of the word of the source hashtag is translated and converted into a hashtag, but not the other words of the source hashtag. For instance, #calgarystampede becomes #stampede de calgary.

The distribution in these categories of the 373 pairs of hashtags where at least one was a multiword hashtag is shown in parentheses in the previous list.

We may only surmise what leads a translator to forego the use of a pound sign and refrain from promoting a word to a hashtag in situations labeled **nil** above. One explanation is that the length of tweets is limited to 140 characters and adding a few pound signs in (the traditionally longer) French text would exceed this limit. However, our observations show that this is improbable: the French text is not that long. More likely, the English hashtag refers to a topic label already used elsewhere in the Twittersphere and its translation would be moot, akin to the translation of a proper noun or monolingual term, hence the existence of examples falling into the **as-is** category described above. Another explanation is that the translator could have forgotten the hashtag or found it irrelevant, maybe hinting at incomplete translation directives and standards in the domain of social media.

As for the **part** category, the examples observed render the hashtags almost useless, sometimes misleading, in the language where the partial tagging occurred. For instance, when #missingchildren becomes #enfants (children), important meaning is lost. Again, this problem could be mitigated by implementing specific translation directives or tools for language professionals.

5. Tweet translation system variants, rationales and results

5.1 Corpora

We present here a number of statistical machine translation (SMT) systems, based on the analysis of the hashtags and tweets we have conducted in the previous sections, putting into (hopefully) good use some of the observations made earlier. To conduct such experiments, we used three non-overlapping corpora, to train the SMT engine, tune its parameters and test it. Here are their respective sources:

- **train**: 2M sentence pairs from Hansard parliamentary debates, 370k sentence pairs from the Canadian website on public safety, 362k sentence pairs from the **url** corpus described in (Gotti et al., 2013) and **twitter-train**
- **tune**: the **twitter-tune** corpus
- **test:** the **twitter-test** corpus

The statistics for these corpora are shown in Table 7. For the English corpora, the percentage of OOV tokens is 2.7% for test and 2.0% for tune. This reflects the fact that **twitter-test** is made from the most recent tweets in each

feed. It therefore constitutes a realistically "hard" corpus. The figures are similar in French.

Corpus	Nb sents	Nb tokens	Nb types
train.en	2737596	34523138	114076
train.fr	2737596	39733914	131069
test.en	2400	38829	5792
test.fr	2400	44132	6240
tune.en	758	12358	3265
tune.fr	758	13964	3446

Table 7: Statistics for the corpora used in this study

5.2 Baseline system and evaluation metrics

r-none is the name of the baseline system in this article. It consists of preprocessing and post-processing steps surrounding a call to the Moses decoder (Koehn et al., 2007). A 5-gram language model was used with Kneser-Ney discounting, trained by the SRILM package (Stolcke, 2002). The "none" in the name **r-none** stands for the fact that no particular mechanism is used to handle hashtags.

Moses was trained, tuned and tested with the corpora described in the previous section, using the default parameters. Java preprocessing includes the modifications mentioned in Section 2. On top of that, the preprocessing module splits each tweet into its constituting sentences and feeds each one separately to Moses, and then post processing reassembles the translation.

Typically, an SMT pipeline is evaluated in terms of BLEU score (Papineni et al., 2002), ranging from 0 (gibberish) to 100 (perfect), by comparing the SMT output to the reference. Word-error rate (WER) is also used. Since this article is interested with hashtags and their translation, we propose additional metrics to measure the SMT's performance when translating them.

We are interested in the recall/precision for the hashtags produced by a translation system, in order to isolate its performance with respect to this element only. We call these metrics hash-R, hash-P, hash-F for hashtag recall, precision and F-measure, respectively.

Another appealing quality metric is the BLEU translation quality metric, but this time for the reference and candidate texts stripped of their pound signs. We call this adapted metric $BLEU_{nohash}$. For a given translation system, it stands to reason that if $BLEU_{nohash} > BLEU$, then the precision of hashtag production in translations is poor, and that spurious hashtags in the translation hurt the performance of the system, where a simple word would have sufficed. An alternate explanation is that the reference (human) translation may not contain enough hashtags to account for those in the source text, which is likely in light of what was discussed in Section 4. The situation where $BLEU_{nohash} < BLEU$ cannot occur.

The metrics for **r-none** with respect to these metrics is shown in Table 8.

Metric	en → fr	fr → en
WER %	48	46
BLEU %	36.61	34.11
$BLEU_{nohash}$ %	37.07	34.86
hash-R %	68	62
hash-P %	63	71
hash-F %	65	66

Table 8: Translation performance for **r-none** (baseline)

Overall, the scores indicate relatively good translations. The translation to French is (unexpectedly) better than the translation into English. The perplexities of the language model on **twitter-test** are 121.3 for French and 252.2 for English. The gain in BLEU scores between the original corpus (BLEU) and the corpora without pound signs ($BLEU_{nohash}$) indicates that we presumably could do better in restoring hashtags in the target text, at least according to the human reference. Nevertheless, a hash-F at 65% is promising, and shows that a train corpus containing as little as 2400 tweet pairs suffices to make the SMT engine able to translate roughly two thirds of source hashtags, which corroborates the distribution observed in Figure 3: some hashtags are indeed very frequent and apparently distributed over train, tune and test material.

5.3 Epilogues and prologues

In section 3.2, we showed that there were distinct parts of a tweet that do not behave like normal text. We decided to treat the prologues and epilogues as distinct units, for the system **r-epipro**. The results are shown in Table 9, along with the difference (in parentheses) in BLEU scores from those reported for **r-none**. The value of this strategy is unequivocal: it does help the translation (albeit not very significantly). We therefore use this strategy in the results that follow. We think it is a sound way of segmenting sentences, and may help in the future if the prologue and epilogue are to be treated in a specific way.

Metric	en → fr	fr → en
WER %	48	46
BLEU %	36.81 (+0.20)	34.46 (+0.35)
$BLEU_{nohash}$ %	37.28 (+0.21)	35.23 (+0.37)
hash-R %	67	62
hash-P %	63	71
hash-F %	65	66

Table 9: Translation performance for **r-epipro**

5.4 Simple lattice input for hashtag translation

Section 3.3 showed that about 80% of hashtags found in the training corpus **twitter-all** have a counterpart in the traditional English (or French) vocabulary, as long as they are stripped of their pound signs (`#`). A hashtag like `#health` can therefore be submitted as `health` to the translation engine, in the hope that it will help translate the word. Since the training corpus may also contain the translation for `#health` as is, it would be best to provide both alternatives as input to the decoder. We perform this by feeding *lattices* to Moses, using its built-in ability to

treat such lattice-encoded alternative inputs. See (Dyer et al., 2008) for more on the lattice system.

The results of this system, **r-lattice**, is shown in Table 10, along with the difference in BLEU scores from the **r-epipro** system presented in the previous section. For both translation directions, the results show a sharp decrease in BLEU, but a gain in $BLEU_{nohash}$. This is consistent with a significant preference of the decoder for the stripped version of each hashtag (i.e. `health` instead of `#health`), presumably driven by the language model. This preference results in fewer actual hashtags being used as input and translated (into hashtags), and more stripped text versions being translated. The translation of plain text sentences as input is simpler for the decoder, trained essentially on just such sentences (hashtags account for a small fraction of training material – see Section 2). In turn, this produces fewer hashtags (hash-R is only 21% for French, compared with 62% for **r-epipro),** and a lower BLEU score. The $BLEU_{nohash}$ score is higher, partly because the decoder translated plain-text sentences more accurately.

Metric	en → fr	fr → en
WER	49	47
BLEU	35.54 (–1.27)	32.50 (–1.96)
$BLEU_{nohash}$	38.56 (+1.28)	36.33 (+1.10)
hash-R	25	21
hash-P	63	72
hash-F	36	32

Table 10: Translation performance for **r-lattice**

5.5 Complete lattice input for hashtags

The previous strategy can be refined to take advantage of the segmentation algorithm we proposed in Section 3.3. Whenever the hashtag stripped of its pound sign cannot be found in the vocabulary, all its segmentations are submitted to the translation engine as a lattice. We rely on the lattice decoding to pick the most probable path in the lattice. All paths receive the same probability at this point, although this could be refined.

A French hashtag like `#nouvelleécosse`, for instance can be split in (among other splits) `[nouvelle, écosse]` or in `[no, uv, el, le, écosse]`. We call the resulting system **r-fulllattice**. The results are shown in Table 11. Differences in BLEU scores from **r-epipro** are reported in parentheses.

Metric	en → fr	fr → en
WER	49	48
BLEU	35.22 (–1.59)	32.42 (–2.04)
$BLEU_{nohash}$	38.23 (+0.95)	36.22 (+0.99)
hash-R	23	19
hash-P	74	87
hash-F	35	31

Table 11: Translation performance for **r-fulllattice**

The results improve upon the **r-epipro** system according to $BLEU_{nohash}$, but not as much as **r-lattice** (previous section). When considering BLEU, the performance is unanimously worse than **r-epipro** and **r-lattice**. The only redeeming quality of the system is its precision in hashtag translation, which is significantly better (roughly +10%) than **r-lattice**, and better than **r-epipro** too. This is not suprising: using such a convoluted way to translate hashtags does not resemble the way humans transpose hashtags when translating, at least according to the observations made earlier on multi-word hashtags. Nevertheless, the approach has the merit of pre-processing hashtags that would have otherwise perturbed the SMT engine, by reducing the OOV rate of input text, as shown by the $BLEU_{nohash}$ score.

6. Conclusion

This study aimed at providing a quantitative analysis of hashtags as observed in the setting of Canadian government agencies' Twitter feeds, and with a view to translating these posts.

Unsurprisingly, hashtags are ubiquitous in tweets. About 40% of them occur outside of the tweet's main text: in what we called epilogues and prologues. Taking advantage of this fact allowed us to refine our tweets' sentence-splitting algorithm, which yielded a direct benefit when implemented in an SMT pipeline. It would be interesting to analyse the form (the "syntax") of these epilogues and prologues. This could provide rules to translate them, without resorting to a full-fledged SMT engine as we did here.

The observed distribution of hashtags hints at their nature: a few of them are used ubiquitously to link the post to recurring themes in Canadian Twitter feeds (such as Canada, politics, health, etc.) while the rarer ones reveal the temporal nature of tweets, sometimes posted to promote a single event or topic. Had we sampled a more varied and greater volume of messages, we might have found fewer happaxes and a greater variety of recurring tags, effectively "flattening" the curve shown in Figure 2.

In the context of translation, hashtags pose a challenge because they are often OOV tokens. We showed that training an SMT engine with a few thousand pairs of tweets suffices to obtain an F-measure of about 66% with respect to hashtag translation, owing to the observed recurrent nature of some of them. Attempting to improve translation quality and retaining hashtag output precision clearly appear as mutually exclusive goals. Indeed, it is feasible to present the SMT engine with the original hashtag and some alternatives (the "dehashed" version, or a segmented version of multiword hashtags), but translation quality improves at the cost of the recall of hashtags, literally lost in translation as the language and translation models favour their non-hashtag equivalents. It is noteworthy that humans also find the problem difficult, as seen from the reference translations mined from Twitter.

Ultimately, it may prove that reconciling these objectives can only be achieved by using another technique. One

such technique is to identify and "dehash" the tags in the source language, translate the text, and then attempt to find equivalent words in the target text in order to promote them to hashtags. This promotion could be guided by the source language hashtags, by the detection of the topic of the source tweet and by the candidate tokens produced by the SMT engine.

While this is beyond the scope of the present study, we would like to hope that the resources and their analysis presented here may offer some insights in models of tweet translation and, more generally, in the processing of tweets and their hashtags.

7. 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 and the anonymous reviewers for their insightful comments.

8. References

Dyer, C., Muresan, S., and Resnik, P. (2008). Generalizing Word Lattice Translation.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., and Smith, N.A. (2011). Part-of-speech Tagging for Twitter: Annotation, Features, and Experiments. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 42–47.

Gotti, F., Langlais, P., and Farzindar, A. (2013). Translating Government Agencies' Tweet Feeds: Specificities, Problems and (a few) Solutions. In Proceedings of the Workshop on Language Analysis in Social Media, (Atlanta, Georgia: Association for Computational Linguistics), pp. 80–89.

Jehl, L.E. (2010). Machine Translation for Twitter. Master's thesis. University of Edinburgh.

Jehl, L., Hieber, F., and Riezler, S. (2012). Twitter Translation Using Translation-based Cross-lingual Retrieval. In Proceedings of the Seventh Workshop on Statistical Machine Translation, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 410–421.

Koehn, P., Hoang, H., Birch, A., Callison-burch, C., Zens, R., Aachen, R., Constantin, A., Federico, M., Bertoldi, N., Dyer, C., et al. (2007). Moses: Open source toolkit for statistical machine translation. pp. 177–180.

Ling, W., Xiang, G., Dyer, C., Black, A., and Trancoso, I. (2013). Microblogs as Parallel Corpora. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), (Sofia, Bulgaria: Association for Computational Linguistics), pp. 176–186.

O'Connor, B., Krieger, M., and Ahn, D. (2010). TweetMotif: Exploratory Search and Topic Summarization for Twitter. W. Cohen, S. Gosling, W. Cohen, and S. Gosling, eds. (The AAAI Press), pp. 34–35.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, (Philadelphia, Pennsylvania: Association for Computational Linguistics), pp. 311–318.

Roberts, K., Roach, M.A., Johnson, J., Guthrie, J., and Harabagiu, S.M. (2012). EmpaTweet: Annotating and Detecting Emotions on Twitter. In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), N.C. (Conference Chair), K. Choukri, T. Declerck, M.U. Doğan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, eds. (Istanbul, Turkey: European Language Resources Association (ELRA)), pp. 3806–3813.

Stolcke, A. (2002). SRILM – An extensible language modeling toolkit. In In Proceedings Of The 7th International Conference On Spoken Language Processing (ICSLP 2002), pp. 901–904.