Twitter Translation using Translation-Based Cross-Lingual Retrieval

Laura Jehl and **Felix Hieber** and **Stefan Riezler**
Department of Computational Linguistics
Heidelberg University
69120 Heidelberg, Germany
{jehl,hieber,riezler}@cl.uni-heidelberg.de

Abstract

Microblogging services such as Twitter have become popular media for real-time user-created news reporting. Such communication often happens in parallel in different languages, e.g., microblog posts related to the same events of the Arab spring were written in Arabic and in English. The goal of this paper is to exploit this parallelism in order to eliminate the main bottleneck in automatic Twitter translation, namely the lack of bilingual sentence pairs for training SMT systems. We show that translation-based cross-lingual information retrieval can retrieve microblog messages across languages that are similar enough to be used to train a standard phrase-based SMT pipeline. Our method outperforms other approaches to domain adaptation for SMT such as language model adaptation, meta-parameter tuning, or self-translation.

1 Introduction

Among the various social media platforms, microblogging services such as Twitter[1] have become popular communication tools. This is due to the easy accessibility of microblogging platforms via internet or mobile phones, and due to the need for a fast mode of communication that microblogging satisfies: Twitter messages are short (limited to 140 characters) and simultaneous (due to frequent updates by prolific microbloggers). Twitter users form a social network by "following" the updates of other users, either reciprocal or one-way. The topics discussed in Twitter messages range from private chatter to important real-time witness reports.

Events such as the Arab spring have shown the power and also the shortcomings of this new mode of communication. Microblogging services played a crucial role in quickly spreading the news about important events, furthermore they were useful in helping organizers plan their protest. The fact that news on microblogging platforms is sometimes ahead of newswire is one of the most interesting facets of this new medium. However, while Twitter messaging is happening in multiple languages, most networks of "friends" and "followers" are monolingual and only about 40% of all messages are in English[2]. One solution to sharing news quickly and internationally was crowdsourcing manual translations, for example at Meedan[3], a nonprofit organization built to share news and opinion between the Arabic and English speaking world, by translating articles and blogs, using machine translation and human expert corrections.

The goal of our research is to automate this translation process, with a further aim of providing rapid crosslingual data access for downstream applications. The automated translation of microblogging messages is facing two main problems. First, there are no bilingual sentence pair data from microblogging domains available. Second, the colloquial, non-standard language of many microblogging messages makes it very difficult to adapt a machine translation system trained on any of the available bilingual resources such as transcriptions from political organizations or news text.

The approach presented in this paper aims to exploit the fact that microblogging often happens in

[1] http://twitter.com/

[2] http://semiocast.com/publications/2011_11_24_Arabic_highest_growth_on_Twitter

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

Proceedings of the 7th Workshop on Statistical Machine Translation, pages 410–421,
Montréal, Canada, June 7-8, 2012. ©2012 Association for Computational Linguistics

parallel in different languages, e.g., microblog posts related to the same events of the Arab spring were published in parallel in Arabic and in English. The central idea is to crawl a large set of topically related Arabic and English microblogging messages, and use Arabic microblog messages as search queries in a cross-lingual information retrieval (CLIR) setup. We use the probabilistic translation-based retrieval technique of Xu et al. (2001) that naturally integrates translation tables for cross-lingual retrieval. The retrieval results are then used as input to a standard SMT pipeline to train translation models, starting from unsupervised induction of word alignments (Och and Ney, 2000) to phrase-extraction (Och and Ney, 2004) and phrase-based decoding (Koehn et al., 2007). We investigate several filtering techniques for retrieval and phrase extraction (Munteanu and Marcu, 2006; Snover et al., 2008) and find a straightforward application of phrase extraction from symmetrized alignments to be optimal. Furthermore, we compare our approach to related domain adaptation techniques for SMT and find our approach to yield large improvements over all related techniques.

Finally, a side-product of our research is a corpus of around 1,000 Arabic Twitter messages with 3 manual English translations each, which were created using crowdsourcing techniques. This corpus is used for development and testing in our experiments.

2 Related Work

SMT for user-generated noisy data has been pioneered at the 2011 Workshop on Statistical Machine Translation that featured a translation task of Haitian Creole emergency SMS messages[4]. This task is very similar to the problem of Twitter translation since SMS contain noisy, abbreviated language. The research papers related to the featured translation task deploy several approaches to domain adaptation, including crowdsourcing (Hu et al., 2011) or extraction of parallel sentences from comparable data (Hewavitharana et al., 2011).

The use of crowdsourcing to evaluate machine translation and to build development sets was pioneered by Callison-Burch (2009) and Zaidan and Callison-Burch (2009). Crowdsourcing has its limits when it comes to generating parallel training data on the scale of millions of parallel sentences. In our work, we use crowdsourcing via Amazon Mechanical Turk[5] to create a development and test corpus that includes 3 English translations for each of around 1,000 Arabic microblog messages.

There is a substantial amount of previous work on extracting parallel sentences from comparable data such as newswire text (Fung and Cheung, 2004; Munteanu and Marcu, 2005; Tillmann and ming Xu, 2009) and on finding parallel phrases in non-parallel sentences (Munteanu and Marcu, 2006; Quirk et al., 2007; Cettolo et al., 2010; Vogel and Hewavitharana, 2011). The approach that is closest to our work is that of Munteanu and Marcu (2006): They use standard information retrieval together with simple word-based translation for CLIR, and extract phrases from the retrieval results using a clean bilingual lexicon and an averaging filter. In this approach, filtering and cleaning techniques in alignment and phrase extraction have to compensate for low-quality retrieval results. In our approach, the focus is on high-quality retrieval.

As our experimental results show, the main improvement of our technique is a decrease in out-of-vocabulary (OOV) rate at an increase of the percentage of correctly translated unigrams and bigrams. Similar work on solving domain adaptation for SMT by mining unseen words has been presented by Snover et al. (2008) and Daumé and Jagarlamudi (2011). Both approaches show improvements by adding new phrase tables; however, both approaches rely on techniques that require larger comparable texts for mining unseen words. Since in our case documents are very short (they consist of 140 character sequences), these techniques are not applicable. However, the advantage of the fact that microblog messages resemble sentences is that we can apply standard word- and phrase-alignment techniques directly to the retrieval results.

Further approaches to domain adaptation for SMT include adaptation using in-domain language models (Bertoldi and Federico, 2009), meta-parameter tuning on in-domain development sets (Koehn and Schroeder, 2007), or translation model adaptation

[4] http://www.statmt.org/wmt11/featured-translation-task.html

[5] http://www.turk.com

using self-translations of in-domain source language texts (Ueffing et al., 2007). In our experiments we compare our approach to these domain adaptation techniques.

3 Cross-Lingual Retrieval via Statistical Translation

3.1 Retrieval Model

In our approach, comparable candidates for domain adaptation are selected via cross-lingual retrieval. In a probabilistic retrieval framework, we estimate the probability of a relevant document microblog message D given a query microblog message Q, $P(D|Q)$. Following Bayes rule, this can be simplified to ranking documents according to the likelihood $P(Q|D)$ if we assume a uniform prior over documents.

$$score(Q, D) = P(D|Q) = \frac{P(D)P(Q|D)}{P(Q)} \tag{1}$$

Our model is defined as follows:

$$score(Q, D) = P(Q|D) = \prod_{q \in Q} P(q|D) \tag{2}$$

$$P(q|D) = \lambda \underbrace{P_{mix}(q|D)}_{\text{mixture model}} + (1 - \lambda) \underbrace{P_{ML}(q|C)}_{\text{query collection backoff}} \tag{3}$$

$$P_{mix}(q|D) = \beta \underbrace{\sum_{d \in D} T(q|d) P_{ML}(d|D)}_{\text{translation model}} + (1 - \beta) \underbrace{P_{ML}(q|D)}_{\text{self-translation}} \tag{4}$$

Our retrieval model is related to monolingual retrieval models such as the language-modeling approach of Ponte and Croft (1998) and the monolingual statistical translation approach of Berger and Lafferty (1999). Xu et al. (2001) extend the former approaches to the cross-lingual setting by adding a term translation table. They describe their model in terms of a Hidden Markov Model with two states that generate query terms: First, a *document state* generates terms d in the document language and then translates them into a query term q. Second, a *backoff state* generates query terms q directly in the query language. In the *document state* the probability of emitting q depends on all d that translate to q, according to a translation distribution T. This is estimated by marginalizing out d as $\sum_d T(q|d)P(d|D)$. In the *backoff state* the probability $P_{ML}(q|C)$ of emitting a query term is estimated as the relative frequency of this term within a corpus in the query language. The probability of transitioning into the document state or the backoff state is given by λ and $1 - \lambda$.

We view this model from a smoothing perspective where the backoff state is linearly interpolated with the translation probability using a mixture weight λ to control the weighting between both terms. Furthermore, we expand Xu et al. (2001)'s generative model to incorporate the concept of "self-translation", introduced by Xue et al. (2008) in a monolingual question-answering context: Twitter messages across languages usually share relevant terms such as hashtags, named entities or user mentions. Therefore, we model the event of a query term literally occurring in the document in a separate model that is itself linearly interpolated with a parameter β with the translation model.

We implemented the model based on a Lucene[6] index, which allows efficient storage of term-document and document-term vectors. To minimize retrieval time, we consider only those documents as retrieval candidates where at least one term translates to a query term, according to the translation table T. Stopwords were removed for both queries and documents. Compared to common inverted index retrieval implementations, our model is quite slow since the document-term vectors have to be loaded. However, multi-threading support and batch retrieval on a Hadoop cluster made the model tractable. On the upside, the translation-based model allows greater precision in finding the candidates for comparable microblog messages than simpler approaches that use a combination of *tfidf* matching and n-best query term expansion: The translation-based retrieval exploits all possible alignments between query and document terms which is particularly important for short documents such as microblog messages.

3.2 In-Domain Phrase Extraction

To prepare the extraction of phrases from retrieval results, we conducted cross-lingual retrieval in both directions: retrieving Arabic documents using English microblog messages as queries and vice versa.

[6] `http://lucene.apache.org/core/`

For each run we kept the top N retrieved documents. Each document was then paired with its query to generate pseudo-parallel data.

We tried two approaches for using this data to improve our translations. The first, more restrictive method makes use of the word alignments we obtained from 5.8 million clean parallel training data from the NIST evaluation campaign. The retrieval step generates word-alignments in the direction $D \rightarrow Q$. After retrieval, the reverse alignment for each query-document pair is also generated by using a translation table in the direction $Q \rightarrow D$. An alignment point between a query term q and a document term d is created, iff $T(q|d)$ or $T(d|q)$ exist in the translation tables $D \rightarrow Q$ or $Q \rightarrow D$. Based on these word-alignments, we extract phrases by applying the *grow-diag-final-and* heuristic and using Och and Ney (2004)'s phrase extraction algorithm as implemented in Moses[7] (Koehn et al., 2007). We conducted experiments using different constraints on the number of alignment points required for a pair to be considered as well as the value of N. Our first technique resembles the technique of Munteanu and Marcu (2006) who also perform phrase extraction by combining clean alignment lexica for initial signals with heuristics to smooth alignments for final fragment extraction.

While we obtained some gains using our heuristics, we are aware that our method is severely restricted in that it only learns new words which are in the vicinity of known words. We therefore also tried the bolder approach of treating our data as parallel and running unsupervised word alignment[8] (Och and Ney, 2000) directly on the query-document pairs to obtain new world alignments and build a phrase table. In contrast to previous work (Snover et al., 2008; Daumé and Jagarlamudi, 2011), we can take advantage of the sentence-like character of microblog messages and treat queries and retrieval results similar to sentence aligned data.

For both extraction methods, the standard five translation features from the new phrase table (phrase translation probability and lexical weighting in both directions, phrase penalty) were added to the translation features in Moses. We tried different modes of combining new and original phrase table, namely using either one or using the new phrase table as backoff in case no phrase translation is found in the original phrase table.

al-Gaddafi, al-Qaddhafi, assad, babrain, bahrain, egypt, gadaffi, gaddaffi, gaddafi, Gheddafi, homs, human rights, human-rights, humanrights, libia, libian, libya, libyan, lybia, lybian, lybya, lybyan, manama, Misrata, nabeelrajab, nato, oman, PositiveLibyaTweets, Qaddhafi, sirte, syria, tripoli, tripolis, yemen;

Table 1: Keywords used for Twitter crawl.

4 Data

4.1 Twitter Crawl

We crawled Twitter messages from September 20, 2011 until January 23, 2012 via the Streaming API[9] in keyword-tracking mode, obtaining 25.5M Twitter messages (*tweets*) in various languages. Table 1 shows the list of keywords that were chosen to retrieve microblog messages related to the events of the Arab spring.[10]

In order to separate the microblog message corpus by languages, we applied a Naive Bayes language identifier[11]. This yielded a distribution with the six most common languages (of 52) being Arabic (57%), English (33%), Somali (2%), Spanish (2%), Indonesian (1.5%), German (0.7%). We kept only microblog messages classified as English or Arabic with confidence greater 0.9. Keyword-based crawling creates a strong bias towards the domain of the keywords and it does not guarantee that all microblog messages regarding a certain topic or region are retrieved or that all retrieved messages are related to the Arab Spring and human righs in the middle east. Additionally, retweets artificially in-

[7] http://statmt.org/moses/

[8] http://code.google.com/p/giza-pp/

[9] https://dev.twitter.com/docs/streaming-api/

[10] The Twitter Streaming API allows up to 400 tracking keywords that are matched to uppercase, lowercase and quoted variations of the keywords. Partial matching such as "tripolis" matching "tripoli" as well as Arabic Unicode characters are not supported. We extended our keywords over time by analyzing the crawl, e.g., by introducing spelling variants and hashtags.

[11] Language Detection Library for Java, by Shuyo Nakatani (http://code.google.com/p/language-detection/).

	Arabic	English
tweets + retweets	14,565,513	8,501,788
tweets	6,614,126	5,129,829
avg. retweet/tweet	11.62	7.27
unique users	180,271	865,202
avg. tweets/user	36.6	5.9

Table 2: Twitter corpus statistics

flate the size of the data, although there are no new terms added. Therefore, we removed all duplicate retweets that did not introduce additional terms to the original tweet. Table 2 explains the shrinkage of the dataset after removing retweets - compared to English users, a smaller number of Arabic users produced a much larger number of retweets. Interestingly, 56,087 users tweet a substantial amount in both languages. This suggests that users spread messages simultaneously in Arabic and English.

4.2 Creating a Small Parallel Twitter Corpus using Crowdsourcing

For the evaluation of our method, a small amount of parallel in-domain data was required. Since there are no corpora of translated microblog messages, we decided to use Amazon Mechanical Turk[12] to create our own evaluation set, following the exploratory work of Zaidan and Callison-Burch (2011b). We randomly selected 2,000 Arabic microblog messages. Hashtags, user mentions and URLs were removed from each microblog message beforehand, because they do not need to be translated and would just artificially inflate scores at test time. The microblog messages were then manually cleaned and pruned. We discarded messages which contained almost no text or large portions of other languages and removed remaining Twitter markup. In the end, 1,022 microblog messages were used in the Mechanical Turk task. We split the data into batches of ten sentences which comprised one HIT (human intelligence task). Each HIT had to be completed by three workers. In order to have some control over translation quality, we inserted one control sentence per HIT, taken from the LDC-GALE Phase 1 Arabic Blog Parallel Text. Turkers were rewarded 10 cents per translation. Following Zaidan and Callison-Burch (2011b), all Arabic sentences were converted into images in order to prevent turkers from pasting them into online machine translation engines. Our final corpus consists of 1,022 translated microblog messages with three translations each. An example containing translations for one of the sentences which we inserted for quality checking purposes, along with the reference translation, is given in table 3. It can be seen that translators sometimes made grammar mistakes or odd word choices. They also tended to omit punctuation marks. However, translations also contained reasonable translation alternatives (such as “gathered” or “collected”). We also asked translators to insert an “unknown” token whenever they were unable to translate a word. Our HIT setup did not allow workers to skip a sentence, forcing them to complete an entire batch. In order to account for translation variants we decided to use all three translations obtained via Mechanical Turk as multiple references instead of just keeping the top translation. We randomly split our small parallel corpus, using half of the microblog messages for development and half for testing.

4.3 Preprocessing

Besides removal of Twitter markup, several additional preprocessing steps such as digit normalization were applied to the data. We also decided to apply the Buckwalter Arabic transliteration scheme[13] to avoid encoding difficulties. Habash and Sadat (2006) have shown that tokenization is helpful for translating Arabic. We therefore decided to apply a more involved tokenization scheme than simple whitespace splitting to our data. As the retrieval relies on translation tables, all data need to be tokenized the same way. We are aware of the MADA+TOKAN Arabic morphological analyzer and tokenizer (Habash and Rambow, 2005), however, this toolkit produces very in-depth analyses of the data and thus led to difficulties when we tried to scale it to millions of sentences/microblog messages. That is why we only used MADA for transliteration and chose to implement the simpler approach by Lee et al. (2003) for tokenization. This approach only requires a small set of annotated data to obtain a list of prefixes and suffixes and uses n-

[12] `http://www.turk.com`

[13] `http://www.qamus.org/transliteration.htm`

REFERENCE	breaking the silence, a campaign group made up of israeli soldiers, gathered anonymous accounts from 26 soldiers.
TRANSLATION1	and breaking silence is a group of israeli soldiers that had unknown statistics from 26 soldiers israeli
TRANSLATION2	breaking the silence by a group of israeli soldiers who gathered unidentified statistics from 26 israeli soldier.
TRANSLATION3	breaking the silence is a group of israeli soldiers that collected unknown statistics of 26 israeli soldiers

Table 3: Example turker translations.

gram-models to determine the most likely *prefix**-*stem-suffix** split of a word.[14]

5 Twitter Translation Experiments

We conducted a series of experiments to evaluate our strategy of using CLIR and phrase-extraction to extract comparable data in the Twitter domain. We also explored more standard ways of domain adaptation such as using English microblog messages to build an in-domain language model, or generating synthetic bilingual corpora from monolingual data.

All experiments were conducted using the Moses machine translation system[15] (Koehn et al., 2007) with standard settings. Language models were built using the SRILM toolkit[16] (Stolcke, 2002). For all experiments, we report lowercased BLEU-4 scores (Papineni et al., 2001) as calculated by Moses' `multi-bleu` script. For assessing significance, we apply the approximate randomization test (Noreen, 1989; Riezler and Maxwell, 2005). We consider pairwise differing results scoring a p-value < 0.05 as significant.

Our baseline model was trained using 5,823,363 million parallel sentences in Modern Standard Arabic (MSA) (198,500,436 tokens) and English (193,671,201 tokens) from the NIST evaluation campaign. This data contains parallel text from different domains, including UN reports, newsgroups, newswire, broadcast news and weblogs.

5.1 Domain Adaption using Monolingual Resources

As a first step, we used the available in-domain data for a combination of domain adaptation techniques similar to Bertoldi and Federico (2009). There were three different adaptation measures: First, the turker-generated development set was used for optimizing the weights of the decoding meta-parameters, as introduced by Koehn and Schroeder (2007). Second, the English microblog messages in our crawl were used to build an in-domain language model. This adaptation technique was first proposed by Zhao et al. (2004). Third, the Arabic portion of our crawl was used to synthetically generate additional parallel training data. This was accomplished by machine-translating the Arabic microblog messages with the best system after performing the first two adaptation steps. Since decoding is very time-intensive, only 1 million randomly selected Arabic microblog messages were used to generate synthetic parallel data. This new data was then used to train another phrase table. Such self-translation techniques have been introduced by Ueffing et al. (2007). All results were evaluated against a baseline of using only NIST data for translation model, language model and weight optimization.

Our results are shown in table 4. Using an in-domain development set while leaving everything else untouched led to an improvement of approximately 1 BLEU point. Three experiments involving the Twitter language model confirm Bertoldi and Federico (2009)'s findings that the language model was most helpful. The BLEU-score could be improved by 1.5 to 2 points in all experiments. When using an in-domain language model, there was no significant difference between deploying an in-domain or out-of-domain development set. We also compared the effect of using only the in-domain language model to that of adding the in-domain language model as an extra feature while keeping the NIST language model.[17] There was no signif-

[14] The n-gram-model required for tokenization was trained on 5.8 million Modern Standard Arabic sentences from the NIST evaluation campaign. This data had previously been tokenized with the same method, trained to match the Penn Arabic Treebank, v3.

[15] `http://statmt.org/moses/`

[16] `http://www.speech.sri.com/projects/srilm/`

[17] The weights for both language models were optimized along with all other translation feature weights, rather than running an extra optimization step to interpolate between both language models, since Koehn and Schroeder (2007) showed that

Run	Translation Model	Language Model	Dev Set	BLEU %
1	NIST	NIST	NIST	13.90
2	NIST	NIST	Twitter	14.83*
3	NIST	Twitter	NIST	15.98*
4	NIST	Twitter	Twitter	15.68*
5	NIST	Twitter & NIST	Twitter	16.04*
6	self-train	Twitter & NIST	Twitter	15.79*
7	self-train & NIST	Twitter & NIST	Twitter	15.94*

Table 4: Domain adaptation experiments. Asterisks indicate significant improvements over baseline (1).

Run	Twitter Phrases	extraction method	# sentence pairs	# extracted phrases	BLEU %
8	top 3 retrieval results	heuristics	14,855,985	6,508,141	17.04*
9	top 1 retrieval results	GIZA++	5,141,065	54,260,537	18.73**
10	retrieval intersection	GIZA++	3,452,566	29,091,009	18.85**
11	retrieval intersection as backoff	GIZA++	3,452,566	29,091,009	18.93**

Table 5: CLIR domain adaptation experiments. All weights were optimized on the Twitter dev set and used the Twitter and NIST language models. One Asterisk indicates a significant improvement over baseline run (5) from table 4. Two Asterisks indicate a significant improvement over run (8).

icant difference between both runs. However, for further adaptation experiments we used the system with the highest absolute BLEU score. In our case, using synthetically generated data was not helpful, yielding similar results as the language model experiments above. As has been observed before by Bertoldi and Federico (2009), it did not matter whether the synthetic data were used on their own or in addition to the original training data.

5.2 Domain Adaptation using Translation-based CLIR

Meta-parameters $\lambda, \beta \in [0, 1]$ of the retrieval model were tuned in a mate-finding experiment: Mate-finding refers to the task of retrieving the single relevant document for a query. In our case, each source tweet in the crowdsourced development set had exactly one “mate”, namely the crowdsourced translation that was ranked best in a further crowdsourced ranking task. Using the retrieval model described in section 3 we achieved precision@1 scores above 95% in finding the translations of a tweet when λ and β were set to 0.9. We fixed these parameter settings for all following experiments. The translation table was taken from the baseline experiments in table 4. During retrieval, we kept up to 10 highest scoring documents per query.

We first employed heuristic phrase extraction based on the word alignments generated from the NIST data as described above. To avoid learning too much noise, maximum phrase length was restricted to 3 (the default is 7). To evaluate the effects of choosing more restrictive or more lax settings, we ran experiments varying the following configurations:

1. Constraints on alignment points:
   - no constraints,
   - 3+ alignment points in each direction,
   - 3+ alignment points in both directions,
   - 5+ alignment points in both directions.
2. Constraints on retrieval ranking:
   - top 10 results,
   - top 3 results,
   - top 1 results,
   - retrieval intersection (results found in both retrieval directions)

We obtained improvements for all combinations of these configurations. However, we observed that requiring 5 common alignment points was too strict, since few pairs met this constraint. We also noticed that using only the top 3 retrieval results was beneficial to performance, suggesting that more comparable microblog messages were indeed ranked higher.

both strategies yielded the same results.

Using extraction heuristics we gained maximally 1.0 BLEU using the top 3 retrieval results and requiring at least 3 alignment points in both alignment directions (see first line in table 5). However, other configurations produced very similar results.

While heuristics led to small incremental improvements, we achieved a much larger improvement by training a new phrase table from scratch using GIZA++. Again, we restricted maximum phrase length to 3 words. In order to keep phrase table size manageable, we had to restrict retrieval to top-1 results or only use retrieval results in the intersection of retrieval directions. Best results are obtained when combining phrase tables extracted from GIZA++ alignments in the intersection of retrieval results with NIST phrase tables in backoff mode (see last line in table 5).

6 Error Analysis

Our cross-lingual retrieval approach succeeded in finding nearly parallel tweets, confirming our hypothesis that such data actually exists. Examples are given in table 6.

Table 7 shows a more detailed breakdown of our translation scores. First, standard adaptation methods increased n-gram precision, suggesting that using in-domain adaptation data caused the system to choose more suitable words. As expected, there was no reduction in OOVs, since using an in-domain language model and development set does not introduce new vocabulary. Heuristic phrase extraction again produced small improvements in n-gram precision while reducing the number of unknown words. Learning a new phrase table with GIZA++ produced substantial improvements both in OOV-rate and in n-gram precision.

Nevertheless, even the scores of the adapted system are still fairly low and translation quality as judged by inspection of the output can be very poor. This suggests that the language used on Twitter still poses a great challenge, due to its variety of styles as well as the users' tendency to use non-standard spelling and colloquial or dialectal expressions. Our development set contained many different genres, from Qu'ran verses over news headlines to personal chatter. Another difficulty was posed by dialectal Arabic content. To gain an impression of the amount of dialectal content in our data, we used the Arabic Online Commentary Dataset created by Zaidan and Callison-Burch (2011a) to classify our test set. Table 8 shows the distribution of dialects in our test data according to language model probability. This distribution should be viewed with a grain of salt, since the shortness of tweets might cause unreliable results when using a model based on word frequencies for classification. Still, the results suggest that there is a high proportion of dialectal content and spelling variation in our data, causing a large number of OOVs. For example, the preposition في, meaning "in" is often written as فى. Our phrase table trained only on standard Arabic data as well as our extraction heuristic failed to translate this frequently occurring word. Only when retraining a phrase table with GIZA++ did we translate it correctly.

Dialect	# Sentences
Egyptian	141
Levantine	147
Gulf	78
Modern Standard Arabic	145

Table 8: Dialectal content in our test set as classified by the AOC dataset.

Table 9 gives examples of translations generated using different adaptation methods in comparison to the references and the Google translation service to illustrate strengths and weaknesses of our approach. *Example 1* shows a case where unknown words were learned through translation model adaptation. Note that even the Google translator did not recognize the word مسيلات which was transliterated as "Msellat". Zaidan and Callison-Burch (2011a) point out that dialectal variants are often transliterated by Google. Note also, that the unadapted translation erroneously translated the place name "sitra" as "jacket", a mistake which was also made in two of the references and by Google. The same happened to the place name "wadyan", which could also be taken as meaning "and religions". This error was enforced by our preprocessing step incorrectly splitting off the prefix "w" which often carries the meaning "and". In addition to that, the two runs which used translation model adaptation each dropped a part of the input sentence ("in sitra", "firing"). We

ARABIC TWEET	ا ف ب الرئيس الفرنسي يؤكد ان القذافي سيحاكم ويدعوا الليبيين الى الصفح
GOOGLE TRANSLATION	*AFP confirms that the French President Gaddafi Libyans tried to call and forgiveness*
ENGLISH TWEET	french president assures that will be taken to court and tells the libyans to forgive each other
ARABIC TWEET	جهاز تنظيم الاتصالات يقرر زيادة رقم جميع شركات المحمول فى مصر دء ا من الخميس
GOOGLE TRANSLATION	*NTRA decide to increase the number of all mobile operators in Egypt a commencement from Thursday*
ENGLISH TWEET	ntra decide to increase the number of all mobile operators in starting from thursday
ARABIC TWEET	الشهيد امين على احمد يوم يناير عن طريق طلق ناري
GOOGLE TRANSLATION	*Shahid Amin AA Day January through gunshot*
ENGLISH TWEET	martyr amin ali ahmed on jan by gunshot

Table 6: Examples of nearly parallel tweets found by our retrieval method.

Adaptation method	OOV-rate %/absolute	unigram precision %/absolute	bigram precision %/absolute	output length (words)
None	22.56/2216	51.1/5020	20.2/1882	9832
LM and Dev	20.05/2220	51.4/5442	22.1/2227	10595
Retrieval (heuristic)	17.47/1790	53.5/5484	23.6/2299	10246
Retrieval (GIZA++)	4.22/439	56.1/5834	26.1/2575	10395

Table 7: OOV-rate and precision for different adaptation methods.

attribute this to that fact that the phrase table extraction often produced one-to-many alignments when only one alignment point was known. In *Example 2* GIZA++ extraction clearly outperformed heuristic phrase extraction. This example also shows that our method is good at learning proper names. While the first two examples resemble news text, *Example 3* is a more informal message. It is particularly interesting to note that with GIZA++ extraction the term "shabiha" is learned, which is commonly used in Syria to mean "thugs" and specifically refers to armed civilians who assault protesters against Bashir Al-Assad's regime. *Example 4* also shows substantial OOV reduction. However, the term بسنترال الأوبرا ("in Opera Central", the location of Telecom Egypt) is incorrectly translated as "really opera".

7 Conclusion

We presented an approach to translation of microblog messages from the Twitter domain. The main obstacle to state-of-the-art SMT of such data is the complete lack of sentence-parallel training data. We presented a technique that uses translation-based CLIR to find relevant Arabic Twitter messages given English Twitter queries, and applies a standard pipeline for unsupervised training of phrase-based SMT to retrieval results. We found this straightforward technique to outperform more conservative techniques to extract phrases from comparable data and also to outperform techniques using monolingual resources for language model adaptation, metaparameter tuning, or self-translation.

The greatest benefit of our approach is a significant reduction of OOV terms at a simultaneous improvement of correct unigram and bigram translations. Despite this positive net effect, we still find a considerable amount of noise in the automatically extracted phrase tables. Noise reduction by improved pre-processing and by more sophisticated training will be subject to future work. Furthermore, we would like to investigate a tighter integration of CLIR and SMT training by using forced decoding techniques for CLIR and by a integrating a feedback loop into retrieval and training.

Acknowledgments

We would like to thank Julia Ostertag for several iterations of manual error analysis of Arabic translation output.

	EXAMPLE 1
SRC	سترة قوات الشغب تقتحم واديان مترجلة وتطلق مسيلات الدموع
GOOGLE	*Riot troops stormed the jacket and religions foot and launches Msellat tears*
NO ADAPTATION	jacket riot forces storm and religions foot وتطلق مسيلات tears
LM AND DEV	sitra and religions of the foot of the riot forces storm وتطلق مسيلات tears
RETRIEVAL (HEURISTIC)	in sitra riot police storming and religions of tear gas on foot
RETRIEVAL (GIZA++)	the riot police stormed and religions of the foot firing tear gas
REF0	vest riot forces break into wadyan by foot and trough gas tear
REF1	sotra the riot forces enter on foot and shoot tear bombs
REF2	the cover for riot police enters wadian walking and shoot tear bombs
	EXAMPLE 2
SRC	أوباما سيتحدث اليوم عن مقتل العولقى
GOOGLE	*Obama will speak today the death of al-Awlaki*
NO ADAPTATION	أوباما سيتحدث today killed العولقى
LM AND DEV	أوباما سيتحدث friday for the killing of العولقى
RETRIEVAL (HEURISTIC)	أوباما today on the killing of
RETRIEVAL (GIZA++)	obama today on the al awlaki killing
REF0	obama will talk today about the killing of al - awlaki
REF1	obama is talking today about el awlaqi death
REF2	obama will speak today about the killing of al - awlaqi
	EXAMPLE 3
SRC	الشبيحة في حماة يستغيثون :)
GOOGLE	Cbihh in Hama are crying :)
NO ADAPTATION	الشبيحة mired in calling for help :)
LM AND DEV	الشبيحة in hama calling for help :)
RETRIEVAL (HEURISTIC)	inside the protectors of the calling for help :)
RETRIEVAL (GIZA++)	shabiha in hama calling for help :)
REF0	the gangsters in hama are asking for help
REF1	the gangs in hamah are peading :)
REF2	the thugs in hama are calling for help :)
	EXAMPLE 4
SRC	حـريـه :: عاملون بالمصرية للاتصالات يحتجزون رئيس الشركة فى غرفة بسنترال الأوبرا
GOOGLE	*Freedom :: Telecom Egypt workers holding company's president in a room Psontral Opera*
NO ADAPTATION	: : free workers بالمصرية للاتصالات holding company chairman فى بسنترال الأوبرا chamber
LM AND DEV	: : workers free بالمصرية للاتصالات holding company chairman بسنترال الأوبرافى room
RETRIEVAL (HEURISTIC)	free : : afcd بالمصرية hold ceo hostage ppl is the president of the chamber of بسنترال الأوبرا
RETRIEVAL (GIZA++)	egypt : : workers telecom workers are holding the head of the company in the chamber of really opera
REF0	freedom :: workers in the egyptian for communication are holding the company president in a room in the opera central
REF1	freedom , workers in egypt for calls detain the head of the company in a room in opera central
REF2	hurriya :: workers in telecom egypt detaining the president of the company in a room in the opera central

Table 9: Example output using different adaptation methods.

References

Adam Berger and John Lafferty. 1999. Information retrieval as statistical translation. In *Proceedings of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, Berkeley, CA.

Nicola Bertoldi and Marcello Federico. 2009. Domain adaptation for statistical machine translation with monolingual resources. In *Proceedings of the 4th EACL Workshop on Statistical Machine Translation*, Athens, Greece.

Chris Callison-Burch. 2009. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, Singapore.

M. Cettolo, M. Federico, and N. Bertoldi. 2010. Mining parallel fragments from comparable texts. In *Proceedings of the 7th International Workshop on Spoken*

Language Translation, Paris, France.

Hal Daumé and Jagadeesh Jagarlamudi. 2011. Domain adaptation for machine translation by mining unseen words. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT'11)*, Portland, OR.

Pascale Fung and Percy Cheung. 2004. Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP'04)*, Barcelona, Spain.

Nizar Habash and Owen Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL'05)*, Ann Arbor, MI.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference - North American Chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL'06)*, New York, NY.

Sanjika Hewavitharana, Nguyen Bach, Qin Gao, Vamshi Ambati, and Stephan Vogel. 2011. CMU haitian creole-english translation system for WMT 2011. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, Edinburgh, Scotland, UK.

Chang Hu, Philip Resnik, Yakov Kronrod, Vladimir Eidelman, Olivia Buzek, and Benjamin B. Bederson. 2011. The value of monolingual crowdsourcing in a real-world translation scenario: Simulation using haitian creole emergency SMS messages. In *Proceedings of the 6th Workshop on Statistical Machine Translation*, Edinburgh, Scotland, UK.

Philipp Koehn and Josh Schroeder. 2007. Experiments in domain adaptation for statistical machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Birch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, Czech Republic.

Young-Suk Lee, Kishore Papineni, Salim Roukos, Ossama Emam, and Hany Hassan. 2003. Language model based arabic word segmentation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics (ACL'03)*, Sapporo, Japan.

Dragos Stefan Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (COLING-ACL'06)*, Sydney, Australia.

Eric W. Noreen. 1989. *Computer Intensive Methods for Testing Hypotheses. An Introduction*. Wiley, New York.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL'00)*, Hongkong, China.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30(4):417–449.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: a method for automatic evaluation of machine translation. Technical Report IBM Research Division Technical Report, RC22176 (W0190-022), Yorktown Heights, N.Y.

Jay M. Ponte and Bruce W. Croft. 1998. A language modeling approach to information retrieval. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'98)*.

Chris Quirk, Raghavendra Udupa U, and Arul Menezes. 2007. Generative models of noisy translations with applications to parallel fragment extraction. In *Proceedings of MT Summit XI*, Copenhagen , Denmark.

Stefan Riezler and John Maxwell. 2005. On some pitfalls in automatic evaluation and significance testing for MT. In *Proceedings of the ACL-05 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, Ann Arbor, MI.

Matthew Snover, Bonnie Dorr, and Richard Schwartz. 2008. Language and translation model adaptation using comparable corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'08)*, Honolulu, Hawaii.

Andreas Stolcke. 2002. SRILM - an extensible language modeling toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, Denver, CO.

Christoph Tillmann and Jian ming Xu. 2009. A simple sentence-level extraction algorithm for comparable data. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North Ameri-*

can Chapter of the Association for Computational Linguistic (NAACL-HLT'09), Boulder, CO.

Nicola Ueffing, Gholamreza Haffari, and Anoop Sarkar. 2007. Transductive learning for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (ACL'07)*, Prague, Czech Republic.

Stephan Vogel and Sanjika Hewavitharana. 2011. Extracting parallel phrases from comparable data. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, Portland, OR.

Jinxi Xu, Ralph Weischedel, and Chanh Nguyen. 2001. Evaluating a probabilistic model for cross-lingual information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'01)*, New York, NY.

Xiaobing Xue, Jiwoon Jeon, and Bruce Croft. 2008. Retrieval models for question and answer archives. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'08)*, Singapore.

Omar F. Zaidan and Chris Callison-Burch. 2009. Feasibility of human-in-the-loop minimum error rate training. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'09)*, Singapore.

Omar F. Zaidan and Chris Callison-Burch. 2011a. The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, Portland, OR.

Omar F. Zaidan and Chris Callison-Burch. 2011b. Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL'11)*, Portland, OR.

Bing Zhao, Matthias Eck, and Stephan Vogel. 2004. Language model adaptation for statistical machine translation with structured query models. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, Geneva, Switzerland.