Improved Arabic Dialect Classification with Social Media Data

Fei Huang
Facebook Inc.
Menlo Park, CA
feihuang@fb.com

Abstract

Arabic dialect classification has been an important and challenging problem for Arabic language processing, especially for social media text analysis and machine translation. In this paper we propose an approach to improving Arabic dialect classification with semi-supervised learning: multiple classifiers are trained with weakly supervised, strongly supervised, and unsupervised data. Their combination yields significant and consistent improvement on two different test sets. The dialect classification accuracy is improved by 5% over the strongly supervised classifier and 20% over the weakly supervised classifier. Furthermore, when applying the improved dialect classifier to build a Modern Standard Arabic (MSA) language model (LM), the new model size is reduced by 70% while the English-Arabic translation quality is improved by 0.6 BLEU point.

1 Introduction

As more and more users share increasing amount of information on various social media platforms (Facebook, Twitter, etc.), text analysis for social media language is getting more important and challenging. When people share their stories, opinions, post comments or tweets on social media platforms, they frequently use colloquial languages, which are more similar to spoken languages. In addition to typical natural language processing problems, the informal nature of social media languages presents additional challenges, such as frequent spelling errors, improper casing, internet slang, spontaneity, dis-fluency and ungrammatical utterances (Eisenstein, 2014). Dialect classification and dialect-specific processing are extra challenges for languages such as Arabic and Chinese.

Considering Arabic as an example: there are big differences between MSA and various dialectal Arabic: MSA is the standardized and literary variety of Arabic used in writing and in most formal speech.[1] It is widely used in government proceedings, newspapers and product manuals. Many research and linguistic resources for Arabic natural language processing are based on MSA. For example, most existing Arabic-English bilingual data are MSA-English parallel sentences. The dialect Arabic has more varieties: 5 major dialects are spoken in different regions of the Arab world: *Egyptian, Gulf, Iraqi, Levantine* and *Maghrebi* (Zaidan and Callison-Burch, 2011). These dialects differ in morphologies, grammatical cases, vocabularies and verb conjugations. These differences call for dialect-specific processing and modeling when building Arabic automatic speech recognition (ASR) systems or machine translation (MT) systems. Therefore, identification and classification of Arabic text is fundamental for building social media Arabic speech and language processing systems.

In order to build better MT systems between Arabic and English, we first analyze the distribution of different Arabic dialects appearing on a very large scale social media platform, as well as their effect on Arabic-English machine translation. We propose several methods to improve the dialect classification accuracy by training models with distant supervision: a weakly supervised model is trained with data whose labels are automati-

[1] http://en.wikipedia.org/wiki/Modern_Standard_Arabic

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2118–2126,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

cally assigned based on authors' geographical information. A strongly supervised model is trained with manually annotated data. More importantly, semi-supervised learning on large amount of unlabeled data effectively increases the classification accuracy. We also combine different classifiers to achieve even bigger improvement. When evaluated on two test sets, the widely adopted Arabic Online Commentary (AOC) corpus and a test set created from the social media domain (Facebook), our methods demonstrate an absolute 20% improvement over the weakly supervised classifier, and 5% over the strongly supervised classifier. Furthermore, the improved classifier is applied on large amount of Arabic social media text to filter out non-MSA data. An LM trained with the cleaned data is used for English-Arabic (MSA) translation. Compared with the baseline model trained with the unfiltered data, the MSA LM reduces the training data by 85%, model size by 70%, and it brings 0.6 BLEU point (Papineni et al., 2002) gain in MT.

The rest of the paper is organized as follows: in section 2 we review previous research on this topic. In section 3 we analyze the dialect distribution and its impact on social media data translation. We present the problem formulation in section 4. In section 5 we introduce two supervised classifiers trained with weakly and strongly labeled data. We describe different semi-supervised learning methods in section 6, followed by the combination of multiple classifiers in section 7. In section 8 we show the experimental results on dialect classification as well as machine translation. The paper finishes with discussions and conclusion in section 9.

2 Related Work

Previous research on Arabic dialect identification focused on two problems: spoken dialect classification for speech recognition ((Novotney et al., 2011) and (Lei and Hansen, 2011)), and written text dialect classification mostly for machine translation. (Habash and Rambow, 2006), (Habash et al., 2008), (Diab et al., 2010) and (Elfardy and Diab, 2012) developed annotation guidelines and morphology analyzer for Arabic dialect.

(Zaidan and Callison-Burch, 2011) created the AOC data set by extracting reader commentary from online Arabic newspaper forums. The selected Arabic sentences are manually labeled with one of 4 dialect labels with the help of crowd sourcing: *Egyptian*, *Gulf*, *Iraqi* and *Levantine*. A dialect classifier using unigram features is trained from the labeled data. In the BOLT (Broad Operational Language Translation) project, translation from dialectal Arabic (especially Egyptian Arabic) to English is a main problem. (Elfardy and Diab, 2013) uses the same labeled AOC data to generate token-based features and perplexity-based features for sentence level dialect identification between MSA and Egyptian Arabic. (Tillmann et al., 2014) trained feature-rich linear classifier based on linear SVM then evaluated the classification between MSA and Egyptian Arabic, reporting 1.4% improvement. All these experiments are based on the AOC corpus. The characteristics and distribution of the Arabic dialects could be different for online social media data. (Darwish et al., 2014) selected Twitter data and developed models taking consideration of lexical, morphological, and phonological information from different dialects, then classified Egyptian and MSA Arabic tweets. (Cotterell and Callison-Burch, 2014) collected dialect data covering Iraqi and Maghrebi Arabic from Twitter as well.

When translating Arabic dialect into English, (Sawaf, 2010) and (Salloum and Habash, 2011) normalized dialect words into MSA equivalents considering character- and morpheme-level features, then translated the normalized input with MSA Arabic-English MT system. (Zbib et al., 2012) used crowd sourcing to build Levantine-English and Egyptian-English parallel data. Even with small amount of parallel corpora for each dialect, they obtained significant gains (6-7 BLEU pts) over a baseline MSA-English MT system.

3 Social Media Arabic Dialect Distribution and Translation

The population speaking a dialect does not necessarily reflect its popularity on internet and social media. Many factors, such as a country's social-economic development status, internet access and government policy, play important roles. To understand the distribution of Arabic dialects on social media, we select data from the largest social media platform, Facebook. There are around one billion users sharing content in 60+ languages every day. The Arabic content comes from different regions of the Arabic world, representative enough for our analysis. We randomly select 2700

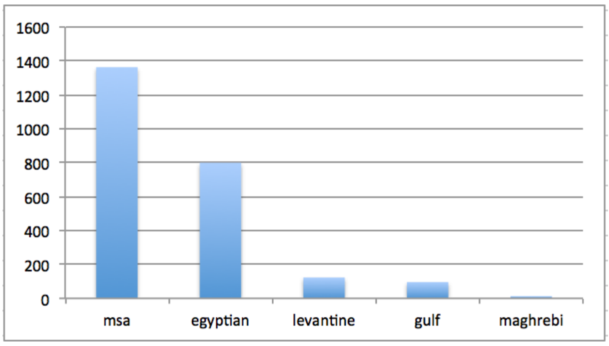


Figure 1: Distribution of various Arabic dialect on the social media platform

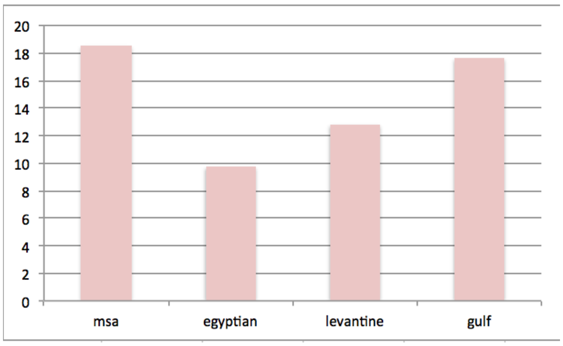


Figure 2: BLEU scores of different Arabic dialects in Arabic-English translation. The MT model is trained with mostly MSA-English parallel data.

sentences from public posts, then ask human annotators to label their dialect types[2]. The result is shown in Figure 1. Not surprisingly, MSA is the most widely used, accounting for 58% of sentences. Besides that, Egyptian Arabic is the most frequent dialect (34%), followed by Levantine and Gulf. Maghrebi is the least frequent. There are other sentences which are not labeled as Arabic dialect, such as classical Arabic, verses from the Quran, foreign words and their transliterations, etc.

We also investigate the effect of different dialects on Arabic-English translation. We ask humans to translate the Arabic sentences into English to create reference translations. We build a phrase-based Arabic-English MT system with 1M sentence pairs selected from MSA Arabic-English parallel corpora (UN corpus, Arabic news corpus, etc.).[3] The training and decoding procedures are similar to those described in (Koehn et al., 2007). More details about the MT system are given in section 8. We group the source Arabic sentences into different subsets based on their dialect labels, then translate them with the MT system. We measure the BLEU score for each subset, as shown in Figure 2. As expected the MSA subset has the highest BLEU scores (18), followed by the Gulf dialect, which is somewhat similar to MSA. The translation of the Egyptian and Levantine dialects is more challenging, with BLEU scores around 10-12, even though they are 40% of the total Arabic data. To improve Arabic-English MT quality, increasing the bilingual data coverage for these two dialects should be most effective, as seen in (Zbib et al., 2012). Because the Maghrebi dialect sample size is too small, we do not report its BLEU score. From these experiments, we further appreciate the importance of accurately identifying Arabic dialect and building dialect-specific translation models.

4 Problem Formulation

In this section we present the general framework of dialect classification. Given a sentence $S = \{w_1, w_2, .., w_l\}$ generated by user u, its dialect class label d^* is determined based on the following functions:

$$d^* = \arg\max_i P(d_i|S, u),$$

where the probability function is defined according to the following exponential model:

$$P(d_i|S, u) = \frac{\exp \sum_k \lambda_k f_k(d_i, \cdot)}{\sum_j \exp \sum_k \lambda_k f_k(d_j, \cdot)}$$

$$d^* = \arg\max_i \sum_k \lambda_k f_k(d_i, \cdot).$$

Here $f_k(d_i, \cdot)$ is the k-th feature function. For example. $f(d_i, u)$ models the likelihood of writing dialect d_i by user u given the user's profile information. $f(d_i, S)$ models the likelihood of generating sentence S with d_i's n-gram language model:

$$\begin{aligned} f(d_i, S) &= \log p(S|d_i) \\ &= \sum_{k=1}^{l} \log p_{d_i}(w_k|w_{k-1}, ..., w_{k-n+1}). \end{aligned}$$

This framework allows the incorporation of rich feature functions such as geographical, lexical,

[2] The data was annotated by a translation service provider under confidentiality agreement.

[3] Because existing Arabic-English bilingual corpora do not include parallel data from social media domain, increasing training data size does not increase the translation quality.

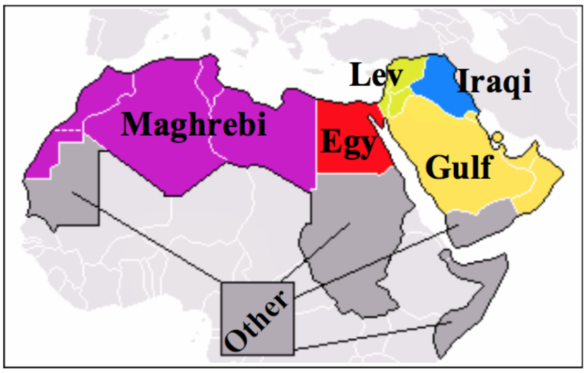


Figure 3: Arabic dialect map, from (Zaidan and Callison-Burch, 2011).

morphological and n-gram information, as seen in previous work ((Zaidan and Callison-Burch, 2011) , (Darwish et al., 2014), (Tillmann et al., 2014) and (Elfardy and Diab, 2013)). However, in this paper we focus on training classifiers with weakly and strongly labeled data, as well as semi-supervised learning methods. So we only choose the geographical and text-based features. Exploration of other features will be reported in another paper.

Previous research (Zaidan and Callison-Burch, 2014) indicated that the unigram model obtains the best accuracy in dialect classification. However, (Tillmann et al., 2014) and (Darwish et al., 2014) exploited more sophisticated text features that lead to better accuracy on selected test set. In our experiments, we find that the unigram model does outperform bigram and trigram models, so we stick to the unigram features.

5 Supervised Learning

5.1 Learning with Weakly Labeled Data

In the chosen social media platform, each user is associated with a unique profile, which includes user-specific information such as the age and gender of the user, the country where s/he is from, etc.. As different Arabic dialects are spoken in different countries, one approach is to classify a post's dialect type based on the author's country, assuming that there is at least a major dialect spoken in each country. This approach is not highly accurate, because the user's country information may be missing or inaccurate; one dialect may be spoken in multiple countries (for example, Egyptian is very popular in different regions of the Arabic world) and multiple dialects may be spoken in the same country; the user can post in MSA instead of dialect Arabic or a mixture of both. However, using data from certain countries as the "approximate" dialect training data, we can train a baseline classifier. As the training data labels are inferred from user profiles instead of manually annotated, such data is called weakly labeled data.

According to the dialect map shown in Fig.3, we group the social media posts into the following 5 dialects according to the author's country:

1. *Egyptian*: Egypt
2. *Gulf*: Saudi Arabia, United Arab Emirate, Qatar, Bahrain, Oman, Yemen
3. *Levantine*: Syrian, Jordan, Palestinian, Lebanese
4. *Iraqi*: Iraq
5. *Maghrebi*: Algeria, Libya, Tunisia, Morocco

Table 1 shows the number of words for each dialect group. Considering the dialect distribution in the social media platform (shown in Figure 1), we focus on the classification of *MSA* (msa) and 3 Arabic dialects: *Egyptian* (egy), *Gulf* (gul) and *Levantine* (lev).

We train an n-gram model for each dialect from the collected data. To train the MSA model, we select sentences from Arabic UN corpus and news collections. All the dialect and MSA models share the same vocabulary, thus perplexity can be compared properly. At classification time, given an input sentence, the classifier computes the perplexity for each dialect type and choose the one with minimum perplexity as the label.

Dialect	Weakly Labeled	Strongly Labeled
egy	22M	0.45M
gul	6M	0.34M
lev	8M	0.45M
msa	27M	1.34M
iraqi	3M	0.01M

Table 1: Corpus size (word count) of weakly and strongly labeled data for supervised learning. The weakly labeled dialect data is from Facebook based on users' country information. The strongly labeled data is manually annotated from the AOC corpus.

5.2 Learning with Strongly Labeled Data

In the AOC corpus, every sentence's dialect type is labeled by human annotators. As these labels are gold labels, the AOC corpus is strongly labeled data. Because of the high cost of manual annotation, the strongly labeled data is much less than the weakly labeled data, but the higher quality makes it possible to train a better classifier. Table 1 shows the corpus size. Although over 50% data is MSA. Egyptian, Gulf and Levantine dialects still have significant presence while the Iraqi dialect has the least labeled data. Such distribution is consistent with what we observed from the social media data. Using these strongly labeled data, we can train a classifier that significant outperforms the weakly supervised classifier.

6 Semi-supervised Learning

6.1 Self-training

Given the small amount of gold labeled data from the AOC corpus and large amount of unlabeled data from the social media platform, a natural combination is semi-supervised learning. In other words, by applying the strongly supervised classifier on the unlabeled data, we can obtain "automatically labeled" dialect data that could further improve the classification accuracy. From the social media platform we select additional Arabic posts with a total of 646M words. The sizes of the newly created dialect corpora are shown in Table 2. Notice that the MSA data accounts for more than 75% of all the labeled data. We train a new classifier with these additional data. As the new labels are only from the original strong classifier, this is self-training.

6.2 Co-Training

Another approach for automatic labeling is co-training (Blum and Mitchell, 1998). With two classifier C_1 and C_2 classifying the same input sentence S with labels l_1 and l_2, S is labeled as l only if $l_1 = l_2 = l$. In other words, a sentence is labeled and used to train a model only when the two classifiers agree. In our experiment we use both the weakly and strongly supervised classifiers to classify the same unlabeled data. Table 2 lists the sizes of the dialect corpora from co-training. Compared with the self-training approach, the co-training method filters out 25% data.

6.3 Data Filtering

Because of domain mismatch, even the strongly supervised classifier does not achieve very high accuracy on the social media test set, thus there is lots of noise in the automatically labeled data. To filter this noise, we only keep the sentences whose minimum perplexity score (corresponding to the winning dialect label) is smaller than any other perplexity score by a margin. Lower perplexity means higher probability of generating the sentence from the dialect model. In other words, sentence S is assigned with label l and used in model re-training if and only if $perp_l(S) < perp_k(S) \times threshold$, for $k \neq l$. The threshold is selected to optimize the classification accuracy on a tuning set. Table 2 also shows the corpora size after filtering. We can see that the filtered dialect is only a quarter of the self-training data. We will compare the three semi-supervised learning methods and evaluate the gains to dialect classification.

Dialect	Self-training	Co-training	Filter
egy	73M	54M	21M
gul	46M	5.1M	2.7M
lev	34M	11.7M	2.5M
msa	493M	406M	139M
All	646M	476M	165M

Table 2: The size of dialect corpora from semi-supervised learning.

7 Classifier Combination

Now we have 3 types of classifiers:

1. The weakly supervised classifier trained with data whose labels are automatically assigned according to author's country;
2. The strongly supervised classifier trained with human labeled data;
3. The semi-supervised classifier trained with automatically classified data, with different data selection methods.

How should we combine them to further improve the classification accuracy?

One approach is data combination: simply adding all the training data together to train a unified n-gram model for each dialect. This experiment is straightforward but the performance is suboptimal because the classifier will be dominated by the model with the most training data, even though its accuracy may not be the best.

The second approach is model combination: we compute the model scores of the weakly supervised (w), strongly supervised (s) and semi-supervised (e) classifiers, then combines them with linear interpolation:

$$p(S|d_i) = \sum_{m=\{w,m,e\}} w_m p_m(S|d_i)$$

As the dialect n-gram perplexity is computed separately, the model weights w_m can be tuned. In our experiments we optimize them with a tuning set from all the dialects.

8 Experiment Results

8.1 Dialect Classification

We already described the training data for supervised and semi-supervised classifiers in previous sections. In this section we will compare their dialect classification accuracies. We select two test sets: 9.5K sentences from the AOC corpus as the AOC test set and 2.3K sentences from the Facebook data set as the FB test set[4]. Both test sets have the dialect of each sentence labeled by human. The accuracy is computed as the percentage of sentences whose classified label is the same as the human label. 90% of the AOC labeled data are used for training the strongly supervised classifier, and the remaining 10% data containing 9.5K sentences is for evaluation. We also keep 200 sentences from the AOC corpus as the development set to tune the model combination parameters.

Model	AOC	FB
weakly supervised	68.4%	48.5%
strongly supervised	83.4%	63.1%
semi-supervised	86.2%	67.7%
combination	87.8%	68.2%

Table 3: Arabic dialect classification accuracies with the weakly and strongly supervised classifiers, as well as the semi-supervised model.

In Table 3 we show the overall classification accuracies of different models on both test sets. Notice that the weakly supervised classifier trained with 68M words obtains 68% accuracy on the AOC test set and 48% on the FB test set (row 1), which is not much higher. However, considering this classifier is trained without any human labeled dialect data, the performance is expected and can be improved with better training data and models. The strongly supervised classifier (row 2), which is trained with much less human labeled data (only 2.6M words), outperforms the weak classifier by 15%. Such a difference is consistently observed in both test tests. This confirms the significant benefits from the gold labeled data.

We apply the strong classifier to large amount of unlabeled data, and train several semi-supervised classifiers with these automatically labeled data. The best result is obtained with the co-training strategy, which brings significant improvement over the strongly supervised model: 2.8-4.6% (row 3), as the label noise is effectively reduced among the agreed labels from two supervised classifiers. Finally, combining all three classifiers (row 1, 2 and 3) with model combination achieves the best result: about 5% improvement of the strong baseline and 20% over the weak baseline. These results demonstrate the effectiveness of combining labeled and unlabeled data obtained from social media platform.

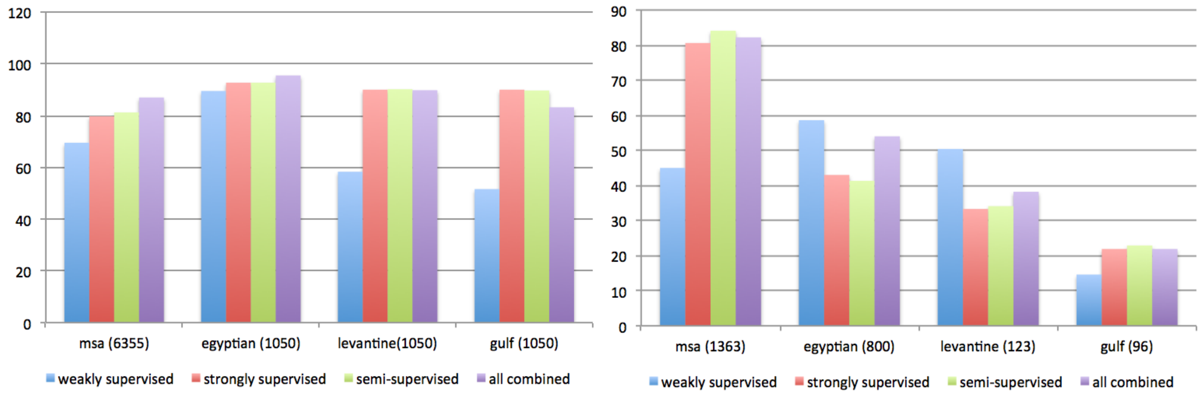
Model	AOC	FB
strongly supervised baseline	83.4%	63.1%
self-training	84.4%	65.5%
co-training	86.2%	67.7%
data filtering	85.2%	64.8%
model interpolation	87.8%	68.2%
data concatenation	82.1%	67.4%

Table 4: Comparison of semi-supervised learning and combination methods.

With semi-supervised learning, we evaluate three data selection methods: self-training, co-training and data filtering. The results are shown in Table 4. Compared with the strong classifier baseline, the self-training method improves by 1% - 2.4%, the co-training method improves by 2.8-4.6%, and the data filtering method improves by 1.7-1.8%. The co-training method is the most effective for both test sets because the information are from two independent classifiers. Data filtering is more effective for the AOC test set (which has the same domain as the baseline model) but less so for the FB test set because valuable in-domain data are filtered out.

In the same table we also compare the results from model and data combination: one from the semi-supervised co-training and the other from the strongly supervised learning. On the AOC test set, the data concatenation method is significantly

[4] The FB test set is available for download at https://www.facebook.com/groups/2419174607/10153205046974608/.



(a) Result on the AOC test set.

(b) Result on the Facebook test set.

Figure 4: Classification precisions by dialect. The number in parenthesis is the number of sentences from each dialect.

worse than the model interpolation method. Its accuracy is even lower than that of the supervised classifier (82.1% vs. 83.4%). However the gap is much smaller on the FB test set. The automatically labeled data is much more than the human labeled data, thus it dominates the combined training data set, which is not a good match for the AOC test data, but is more relevant to the FB test data. In both cases, the model combination obtains better classification accuracies, where the supervised model is assigned higher weights (0.9) and the semi-supervised model is used for smoothing, therefore the combined model is able to improve over the strong classifier.

We further analyze the classification precision for each type of dialect on both test sets in Figure 4. Figure 4a shows the result on the AOC test set. The number after the dialect type (in the parenthesis) is the number of sentences from that dialect. Precisions increase from the weakly supervised to the strongly supervised to the semi-supervised classifier, and the combined classifier generally outperforms all three classifiers, except for the Gulf dialect. However, considering the smaller percentage of the Gulf dialect, we still observe significant improvement overall. Figure 4b shows the result on the FB test set, where the MSA and Egyptian dialects are much more frequent than the Levantine and Gulf dialects. Improving classification on the MSA and Egyptian dialect (especially MSA) will be very helpful. We notice that the supervised classifier improves over the unsupervised classifier by a large margin on the MSA and Gulf dialects, but performs worse on the Egyptian and Levantine dialects. This is different from the result in the AOC test set, where the supervised classifier consistently improves over the unsupervised classifier. One reason is that in the AOC test set, the training and test data are from the same corpus, thus the supervised training from in-domain data is very effective. For the FB test set, the strongly labeled data and the test data mismatch in genre and topics. The automatically labeled data is less similar to the dialect test set, thus it is less effective for the Egyptian and Levantine dialects. This further confirms the necessity of combining information from multiple sources. The combined classifier performs significantly better for the MSA and Gulf dialect, but slightly worse for the Egyptian and Levantine dialects. The overall result is still positive.

We also compare our approach with other dialect classification methods on the AOC corpus, which is commonly used so the results are comparable. Most previous work focus on the classification of MSA vs. EGY dialect, and report the accuracies from 85.3% (Elfardy and Diab, 2013), 87.9% (Zaidan and Callison-Burch, 2014) to 89.1% (Tillmann et al., 2014), adding morphological features, using word-based unigram-model and linear SVM models. Our MSA vs. EGY dialect classification accuracy is 92.0%, the best known result on this test set. We do not use more sophisticated features; the improvement is just from the mined unlabeled data and the combination of different classifiers. On the FB test set, our strongly supervised classifier is the same as (Zaidan and Callison-Burch, 2014), both using word-based unigram model. We see 5% gain with the combined classifiers.

8.2 Machine Translation

The motivation of this research is to handle challenges from Arabic dialects to improve machine translation quality. For example, using the dialect classifier output one can build dialect-specific Arabic-English MT systems. Given an Arabic sentence, the system first identifies its dialect type, then translates with the corresponding MT system. When building English-to-Arabic (MSA) translation systems for social media translation, the target LM trained from in-domain data is very helpful to improve the translation quality. Considering that the Arabic in-domain data contains lots of dialects, an effective dialect classifier helps filter out dialect Arabic and only keep the MSA to train a cleaner LM.

Because of the limited bilingual resources of dialect Arabic-English, we will focus on English-Arabic MT system first. In this experiment, the training data for the English-Arabic MT system is 1M parallel sentences selected from publicly available Arabic-English bilingual data (LDC, OPUS). Because none of the parallel corpora is for social media translation, we select a subset closer to the social media domain by maximizing the n-gram coverage on the test domain. The development and test sets contain 700 and 892 English sentences, respectively. These sentences are translated into MSA by human translators. We apply the standard SMT system building procedures: pre-processing, automatic word alignment, phrase extraction, parameter tuning with MERT, and decoding with a typical phrase-based decoder similar to (Koehn et al., 2007). The LM is trained with the target side of the parallel data, plus 200M in-domain Arabic sentences.

Using the above combined dialect classifier, we label the dialect type of each sentence in the in-domain data, filter out any non-MSA sentences and re-train the target LM. Again to keep the in-domain data clean, we also apply the threshold-based data filtering. As shown in Table 5, the dialect filtering reduces the LM training data by 85%, which corresponds to 70% less memory footprint. Thanks to the cleaner LM, the translation quality is also improved by 0.6 BLEU point.[5]

	All Arabic Data	Filtered MSA data
number of sentences	200M	30M
memory footprint	23G	6.6G
BLEU score (1-reference)	12.52	13.14

Table 5: Cleaned MSA LM after dialect filtering for English-Arabic(MSA) translation.

9 Discussion and Conclusion

Existing Arabic dialect classification methods solely rely on textural features, be they n-gram language model or morphology/POS-based features. This paper utilizes authors' geographical information to train a weakly supervised dialect classifier. Using the weakly and strongly supervised classifiers to classify and filter unlabeled data leads to several improved semi-supervised classifiers. The combination of all three significantly improves the Arabic dialect classification accuracy on both in-domain and out-of-domain test sets: 20% absolute improvement over the weak baseline and 5% absolute over the strong baseline. After applying the proposed classifier to filter out Arabic dialect data, and building a cleaned MSA LM, we observe 70% model size reduction with 0.6 BLEU point gain in English-Arabic translation quality.

In future work, we would like to explore more user-specific information for dialect classification, apply the classifier for Arabic-to-English MT systems, and extend the approach to a larger family of languages and dialects.

References

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM.

Ryan Cotterell and Chris Callison-Burch. 2014. A multi-dialect, multi-genre corpus of informal written arabic. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may.

Kareem Darwish, Hassan Sajjad, and Hamdy Mubarak. 2014. Verifiably effective arabic dialect identification. In *Proceedings of the 2014 Conference on*

[5] Due to the challenging nature of social media data, and the lack of in-domain training data, the BLEU score is much lower than the one in news translation.

Empirical Methods in Natural Language Processing (EMNLP), pages 1465–1468. Association for Computational Linguistics.

Mona Diab, Nizar Habash, Owen Rambow, Mohamed Altantawy, and Yassine Benajiba. 2010. Colaba: Arabic dialect annotation and processing. In *LREC Workshop on Semitic Language Processing*, pages 66–74.

Jacob Eisenstein. 2014. Identifying regional dialects in online social media.

Heba Elfardy and Mona T Diab. 2012. Simplified guidelines for the creation of large scale dialectal arabic annotations. In *LREC*, pages 371–378.

Heba Elfardy and Mona T Diab. 2013. Sentence level dialect identification in arabic. In *ACL (2)*, pages 456–461.

Nizar Habash and Owen Rambow. 2006. Magead: a morphological analyzer and generator for the arabic dialects. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 681–688. Association for Computational Linguistics.

Nizar Habash, Owen Rambow, Mona Diab, and Reem Kanjawi-Faraj. 2008. Guidelines for annotation of arabic dialectness. In *Proceedings of the LREC Workshop on HLT & NLP within the Arabic world*, pages 49–53.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Yun Lei and John HL Hansen. 2011. Dialect classification via text-independent training and testing for arabic, spanish, and chinese. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(1):85–96.

Scott Novotney, Richard M Schwartz, and Sanjeev Khudanpur. 2011. Unsupervised arabic dialect adaptation with self-training. In *INTERSPEECH*, pages 541–544. Citeseer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Wael Salloum and Nizar Habash. 2011. Dialectal to standard arabic paraphrasing to improve arabic-english statistical machine translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21. Association for Computational Linguistics.

Hassan Sawaf. 2010. Arabic dialect handling in hybrid machine translation. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas*.

Christoph Tillmann, Saab Mansour, and Yaser Al-Onaizan, 2014. *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, chapter Improved Sentence-Level Arabic Dialect Classification, pages 110–119. Association for Computational Linguistics and Dublin City University.

Omar F. Zaidan and Chris Callison-Burch. 2011. 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: Human Language Technologies*, pages 37–41, Portland, Oregon, USA, June. Association for Computational Linguistics.

Omar F Zaidan and Chris Callison-Burch. 2014. Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Rabih Zbib, Erika Malchiodi, Jacob Devlin, David Stallard, Spyros Matsoukas, Richard Schwartz, John Makhoul, Omar F Zaidan, and Chris Callison-Burch. 2012. Machine translation of arabic dialects. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 49–59. Association for Computational Linguistics.