Machine Translation in Mobile Games: Augmenting Social Media Text Normalization with Incentivized Feedback

Nikhil Bojja nbojja@machinezone.com
Arun Nedunchezhian arun@machinezone.com
Pidong Wang pwang@machinezone.com
Machine Zone Inc., Palo Alto, CA, USA

Abstract

Machine Translation across languages is made difficult in the context of Mobile games where slang or ungrammatical language reduces the effectiveness of open domain translation systems. We describe a system here that improves translation systems by normalizing user slang with an active learning system. A crowsourcing system is created by incentivizing players to normalize slang through a game feature that rewards participants with in-game currency rewards. The rewards ensure active participation from players and the feedback is in turn used to train a phrase-based Text Normalization System that is relevant to the domain of the data, thereby improving Machine Translation accuracy.

1 Introduction

Advances in Machine Translation techniques have enabled people from across the globe to communicate with each other beyond language boundaries. Online texts such as news articles can be translated on demand with commercial translation service providers. These providers have reasonable translation accuracy with texts under various domains. The problem of accuracy in Machine Translation is made severe when we target general purpose translation systems on domain specific data, especially when this domain specific data is not very grammatical. Applying domain specific data to re-train and adapt translation systems is a potential solution for this problem. However, it is not easy to obtain Social Media or Mobile Game text in a format that can be used to train translation systems.

For our experiments, we selected *Game of War: Fire Age*, a popular Massively Multiplayer Online Role Playing Game (MMORPG) that is primarily played on mobile devices on a global scale. This game has the ability to let players from around the world communicate in real-time with each other and across languages with the help of an in-built translation module. In this paper, we describe this system and the problem of acquiring data for improving machine translation output in the context of slang-speak in mobile game interactions.

In the following sections we will talk about some of the related work in this field, describe the system, showcase improvements brought in by this system and discuss future possibilities.

2 Previous work

Statistical Machine Translation (Brown et al., 1993) has made it easy for people around the world to access webpages in foreign languages. Its applications help make more information available for those seeking it. Phrase-based Statistical Machine Translation (Koehn et al., 2003) has been a popular choice for building machine translation systems between language pairs. Parallel corpora between source and target languages are used to build phrase level alignment tables, which are then used in conjunction with a language model to generate target translations. This makes the model sensitive to the data that it is trained on and specifically the domain of the data supplied.

When it comes to specific domains like mobile games, players communicate with each other in a highly informal setting. Text generated from such a setting tends to have slang words and chats that are not necessarily structured well grammatically, and could have a lot of misspellings. It is known that should we attempt to apply Machine Translation on texts with a lot of informal slang in them, the translation output is less than optimal (Ling et al., 2013). Attempts have been made in the Machine Translation community to normalize the effect of such slang by using slang dictionaries. Aw et al. (2006) have shown that building a Statistical Machine Translation system just for the purpose of normalizing slang can have an overall improvement in translation quality. Another work (Wang and Ng, 2013; Wang, 2013) has presented a novel text rewriting decoder for slang text normalization that could enhance overall translation accuracy of the system.

3 Normalization system

The translation system in *Game of War: Fire Age* lets players chat with each other in realtime. To make this possible chats from a source language are run through *MZ Transformer*, an ensemble normalization system which employs a combination of slang dictionaries, abbreviation lists, spell checkers and most importantly a phrase based text normalization system. To develop the phrase based text normalization system, we prepared a slang corpus made up of player chats extracted from the Mobile Game logs. The data was noted to contain slang used by players in the game and reflected the informal tone of the domain. The slang corpus was then manually normalized to a grammatical equivalent corpus of sentences. The eventual parallel corpus of slang and normalized sentences served as training data for building a Phrase-based Statistical Text Normalization system using Moses (Koehn et al., 2007).

The resulting system *translated* slang text to grammatical text within the same source language. *MZ Transformer* could now handle the transformation of most of player slang used in the game and convert it to a grammatically better version. This grammatical version was then fed to a hybrid translation system which comprised of an internal cross language translation system and commercial translation service providers [1][2]. The overall quality and readability of output translations obtained was observed to be significantly better. More importantly, the system could now make sense of slang used by players than just delegating them as Out of Vocabulary words (OOV's).

4 The Data problem

The initial parallel text used for creating the prototype in Section 3 was manually created. This is of course expensive and not feasible when we want to build a more robust system with a larger training dataset or similar systems for languages other than English. Various methods have been suggested for accumulating bilingual training data for building Statistical Machine Translation systems for instant messaging systems (Bangalore et al., 2002) or for microblogs (Ling et al., 2013; Xu et al., 2013).

Though the vocabulary of these domains can be assumed to be similar to the language

[1] Microsoft Translator. http://bing.com/translator

[2] Google Translate. http://translate.google.com

used in Mobile games, we noticed that this domain uses a more specific vocabulary tied to in-game actions and events. We also noticed that the slang used in games contained many more abbreviations and variations than that of microblogs. The length of source sentences in Mobile games tended to be smaller than microblog messages such as those from Twitter. On identical sample sizes, Twitter messages averaged 73.51 characters per source sentence compared to 34.43 characters per source sentence in the Mobile game dataset (Wang et al., 2015). The length and perplexity of the Mobile game data is hence contextually limited that indicates sub-domain level differences. It should be noted that the limited data per input sentence further exacerbates the lower translation accuracy problem.

Crowdsourcing techniques could be a good way to obtain parallel data in these cases. Platforms such as Amazon's Mechanical Turk *(mTurk)* could be used to obtain data (Zaidan and Callison-Burch, 2011). Apart from the monetary cost associated with it, getting data in languages other than English came up as an issue with using *mTurk*. Thus it became necessary for us to create a novel system to create the dataset of parallel slang and grammatical data at a low cost.

4.1 Game Economy

Most Multiplayer Role playing strategy games have an in-game economy that is critical to its functioning. *Game of War* too has such an economy with in-game currency on one side, and various items available for sale on the other side. The items available are bought by players to be used in the game. There is a huge variety in the types of items available as well as the quantities in which they are offered. In-game currency is used to monitor the pricing of these items and game designers have the flexibility to offer sales and discounted prices on the items available. Needless to say, these in-game purchases are highly sought after by active players who want to get ahead of their competitors in the game. Control over such a lucrative game economy can be leveraged for our purpose of collecting data needed for training our models.

5 Crowdsourcing System

To solve the data problem, we created a Rewards based Normalization module within the game. In this module, players are presented with slang words or phrases that need to be normalized. Along with each such input, in-game items are presented as rewards in remuneration for normalizing the data. This way we provide an engaging feature within the game where players can earn in-game items in exchange for spending some effort normalizing slang text.

Text is injected into the module based on language and number of unknown slang words in the corpus. Each phrase is presented to multiple players concurrently and normalized outputs are accumulated. To ensure high quality output, we setup a two-step process. One set of players type in normalized versions of input slang words/phrases, and another set of players are shown a multiple-choice style visualization of input slang phrase and candidate normalized phrases with an option for users to choose from the normalized output versions. Automated Quality control is put in place by use of text similarity techniques to remove entries entered by users that are irrelevant to the input word/phrase.

5.1 Task Instructions

The only paragraph of instruction that appears to all players participating in the crowsourced task is: *Select the best correction for the misspelled words and earn rewards. We select the top, most accurate entry submitted by users like you and approve rewards for them. Note that there could be cases where theres no correction needed too.*

Given that the feedback system is connected to an online game with a highly active chat system the users of the system discussed the feature and have evolved into a user-base that agreed upon the right way to do the job. We did put in checks to avoid collusion and have been successful in making the system efficient.

5.2 Creating the Parallel Corpus

Source Phrase	Response Received	Num. Users
yo wasup zack .. i just wakey	Yo, what's up Zack? I just woke up.	**1013**
	Hi, what's up Zack? I just woke up.	327
	Hey, what's up Zack? I just woke up.	133
	What's up Zack? I just woke up.	61
	To what's up Zack? I just woke up.	12
	Yo what's up Zack. I just awoke	3

Table 1: Sample of Data collected

The player base in *Game of War: Fire Age* is numerous enough for us to choose a 1-best hypothesis that has been agreed upon by a multitude of players for a given input sentence. There is of course an option to obtain n-best hypotheses - ranked by number of players agreeing on the same normalized output. Rewards are given out to players at the end of selecting the 1-best hypothesis for inputs. A sample of the data collected per phrase can be seen in Table 1. One can see that the top hypothesis is significantly ahead of the remaining hypotheses which validates the use of a 1-best hypothesis. We note that this trend is consistent with data collected across other remaining input phrases too.

Hence, we now have a feedback loop from players who can help improve the normalization process and in turn improve translation accuracy. Such a feedback loop is a desired feature in every Machine Translation system. The lack of incentives could be attributed to users seldom providing feedback on translation quality in traditional translation systems. Due to the game economy based incentives, we have a feedback loop that is assured to gather feedback in a timely manner from a willing player base.

6 Experiments and Discussions

Using the Rewards based Crowdsourcing system we were able to collect normalized data across languages such as English, French, Spanish, German, Portuguese and Russian. Translation systems augmented with *MZ Transformer* as described in Section 3 were built for each of these languages using the data collected.

To measure the impact of the normalization system on translation quality, a separate held out test set was created with manually translated messages in various language pairs. Each language pair had 1000 samples in the test set. The number of tokens in each of the test sets approximately averaged 6500 in number. The test set for each language pair was built through a random selection of chats from a database.

6.1 Results

The test set for each language pair was translated with a commercial translation provider[3] and translated with a Translation system that gets normalized inputs from *MZ Transformer*. We used the BLEU metric (Papineni et al., 2002) to measure the translation quality. The results are shown in Table 2.

The results show a clear improvement in translation quality for all language pairs when Normalization was used as a pre-step before translation. Manual analysis of the outputs showed

[3] Microsoft Translator: `http://bing.com/translator`

Source Lang.	Target Lang.	w/o Normalization	w/ Normalization
Spanish	English	37.82	**39.77**
English	Spanish	31.29	**32.87**
French	English	46.30	**47.73**
English	French	31.90	**33.19**
German	English	41.02	**43.98**
English	German	26.92	**26.96**
Portuguese	English	50.94	**52.13**
English	Portuguese	38.09	**38.12**
Russian	English	38.64	**40.17**
English	Russian	24.80	**25.43**

Table 2: BLEU score improvement

that even in language pairs where the improvement in BLEU scores was minimal, the readability of the sentences improved greatly with normalization. Normalization targets tokens that tend to have a higher degree of occurrence in player chats. As an example, *lol* in English (laugh out loud) is the most frequently occurring token in the player chat database. However, this does not occur as frequently in the test set. *MZ Transformer* however ensures that *lol* is translated to *mdr* when translating to French. *mdr* (mort de rire) is the equivalent of *lol* in French slang. Readability greatly improves in a player chat session with such translations on high frequency slang words, but such gains don't necessarily translate to BLEU score improvements.

A learning from these results is that an improvement in translation quality correlates with the number of normalization layers and the quantity of training data in *MZ Transformer*. Also, each language seems to have a different degree of slang usage and hence we deduce that perplexity correlates with translation improvement too. Do note that this was only one round of feedback addition to *MZ Transformer's* training data. After collecting some more data we could check for further improvements in translation quality.

We used 10-best hypotheses from the data collection process (Table 1) as an alternate training dataset for the Phrase-based text Normalization system. This system had a lower BLEU score compared to the system trained with 1-best hypotheses. This could be attributed to overfitting because of the high degree of similarity in training hypotheses.

7 Future work

The Mobile game economy and the demand for in-game items from players creates an ideal ecosystem where getting crowdsourced data becomes easy. With the growing popularity of Mobile games around the world, getting data on resource poor languages can be made easy through a crowdsourced ecosystem like this where we have access to native speakers of various languages globally. We have started collecting data in languages such as Bulgarian, Malay, Ukrainian, Slovak among others and hope to build similar normalization systems in these languages.

The system could be further utilized to collect data of any kind, be it text normalization, text translation or even speech transcription. The speed at which crowdsourcing is done could be modulated with the number of rewards announced for each task. This will ensure speedy output from the system should we need data urgently. As the number of players outnumbers the amount of data needed, we can get multiple hypotheses for each input, thus ensuring a high quality crowdsourced output.

References

Aw, A., Zhang, M., Xiao, J., and Su, J. (2006). A phrase-based statistical model for sms text normalization. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions*, COLING-ACL '06, pages 33–40, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bangalore, S., Murdock, V., and Riccardi, G. (2002). Bootstrapping bilingual data using consensus translation for a multilingual instant messaging system. In *COLING*.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Comput. Linguist.*, 19(2):263–311.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (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.

Koehn, P., Och, F. J., and Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 48–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ling, W., Xiang, G., Dyer, C., Black, A., and Trancoso, I. (2013). Microblogs as parallel corpora. In *Proceedings of the 51st Annual Meeting on Association for Computational Linguistics*, ACL '13. Association for Computational Linguistics.

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*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wang, P. (2013). *A Text Rewriting Decoder with Application to Machine Translation*. PhD thesis, National University of Singapore.

Wang, P., Bojja, N., and Kannan, S. (2015). A language detection system for short chats in mobile games. In *Proceedings of the third International Workshop on Natural Language Processing for Social Media*, pages 20–28, Denver, Colorado. Association for Computational Linguistics.

Wang, P. and Ng, H. T. (2013). A beam-search decoder for normalization of social media text with application to machine translation. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 471–481, Atlanta, Georgia. Association for Computational Linguistics.

Xu, W., Ritter, A., and Grishman, R. (2013). *Proceedings of the Sixth Workshop on Building and Using Comparable Corpora*, chapter Gathering and Generating Paraphrases from Twitter with Application to Normalization, pages 121–128. Association for Computational Linguistics.

Zaidan, O. F. and Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1220–1229, Stroudsburg, PA, USA. Association for Computational Linguistics.