Acquisition of Translation Lexicons for Historically Unwritten Languages via Bridging Loanwords

Michael Bloodgood
Department of Computer Science
The College of New Jersey
Ewing, NJ 08628
mbloodgood@tcnj.edu

Benjamin Strauss
Computer Science and Engineering Dept.
The Ohio State University
Columbus, OH 43210
strauss.105@osu.edu

Abstract

With the advent of informal electronic communications such as social media, colloquial languages that were historically unwritten are being written for the first time in heavily code-switched environments. We present a method for inducing portions of translation lexicons through the use of expert knowledge in these settings where there are approximately zero resources available other than a language informant, potentially not even large amounts of monolingual data. We investigate inducing a Moroccan Darija-English translation lexicon via French loanwords bridging into English and find that a useful lexicon is induced for human-assisted translation and statistical machine translation.

1 Introduction

With the explosive growth of informal electronic communications such as email, social media, web comments, etc., colloquial languages that were historically unwritten are starting to be written for the first time. For these languages, there are extremely limited (approximately zero) resources available, not even large amounts of monolingual text data or possibly not even small amounts of monolingual text data. Even when audio resources are available, difficulties arise when converting sound to text (Tratz et al., 2013; Robinson and Gadelii, 2003). Moreover, the text data that can be obtained often has non-standard spellings and substantial code-switching with other traditionally written languages (Tratz et al., 2013).

In this paper we present a method for the acquisition of translation lexicons via loanwords and expert knowledge that requires zero resources of the borrowing language. Many historically unwritten languages borrow from highly resourced languages. Also, it is often feasible to locate a language expert to find out how sounds in these languages would be rendered if they were to be written as many of them are beginning to be written in social media, etc. We thus expect the general method to be applicable for multiple historically unwritten languages. In this paper we investigate inducing a Moroccan Darija-English translation lexicon via borrowed French words. Moroccan Darija is an historically unwritten dialect of Arabic spoken by millions but lacking in standardization and linguistic resources (Tratz et al., 2013). Moroccan Darija is known to borrow many words from French, one of the most highly resourced languages in the world. By mapping Moroccan Darija-French borrowings to their donor French words, we can rapidly create lexical resources for portions of Moroccan Darija vocabulary for which no resources currently exist. For example, we could use one of many bilingual French-English dictionaries to bridge into English and create a Moroccan Darija-English translation lexicon that can be used to assist professional translation of Moroccan Darija into English and to assist with construction of Moroccan Darija-English Machine Translation (MT) systems.

The rest of this paper is structured as follows. Section 2 summarizes related work; section 3 explains our method; section 4 discusses experimental results of applying our method to the case of building a Moroccan Darija-English translation lexicon; and section 5 concludes.

2 Related Work

Translation lexicons are a core resource used for multilingual processing of languages. Manual creation of translation lexicons by lexicographers is

Proceedings of the 10th Workshop on Building and Using Comparable Corpora, pages 21–25,
Vancouver, Canada, August 3, 2017. ©2017 Association for Computational Linguistics

time-consuming and expensive. There are more than 7000 languages in the world, many of which are historically unwritten (Lewis et al., 2015). For a relatively small number of these languages there are extensive resources available that have been manually created. It has been noted by others (Mann and Yarowsky, 2001; Schafer and Yarowsky, 2002) that languages are organized into families and that using cognates between sister languages can help rapidly create translation lexicons for lower-resourced languages. For example, the methods in (Mann and Yarowsky, 2001) are able to detect that English *kilograms* maps to Portuguese *quilogramas* via bridge Spanish *kilogramos*. This general idea has been worked on extensively in the context of cognates detection, with 'cognate' typically re-defined to include loanwords as well as true cognates. The methods use monolingual data at a minimum and many signals such as orthographic similarity, phonetic similarity, contextual similarity, temporal similarity, frequency similarity, burstiness similarity, and topic similarity (Bloodgood and Strauss, 2017; Irvine and Callison-Burch, 2013; Kondrak et al., 2003; Schafer and Yarowsky, 2002; Mann and Yarowsky, 2001). Inducing translations via loanwords was specifically targeted in (Tsvetkov and Dyer, 2015; Tsvetkov et al., 2015). While some of these methods don't require bilingual resources, with the possible exception of small bilingual seed dictionaries, they do at a minimum require monolingual text data in the languages to be modeled and sometimes have specific requirements on the monolingual text data such as having text coming from the same time period for each of the languages being modeled. For colloquial languages that were historically unwritten, but that are now starting to be written with the advent of social media and web comments, there are often extremely limited resources of any type available, not even large amounts of monolingual text data. Moreover, the written data that can be obtained often has non-standard spellings and code-switching with other traditionally written languages. Often the code-switching occurs within words whereby the base is borrowed and the affixes are not borrowed, analogous to the multi-language categories "V" and "N" from (Mericli and Bloodgood, 2012). The data available for historically unwritten languages, and especially the lack thereof, is not suitable for previously developed cognates detection methods that operate as discussed above. In the next section we present a method for translation lexicon induction via loanwords that uses expert knowledge and requires zero resources from the borrowing language other than a language informant.

3 Method

Our method is to take word pronunciations from the donor language we are using and convert them to how they would be rendered in the borrowing language if they were to be borrowed. These are our candidate loanwords. There are three possible cases for a given generated candidate loanword string:

true match string occurs in borrowing language and is a loanword from the donor language;

false match string occurs in borrowing language by coincidence but it's not a loanword from the donor language;

no match string does not occur in borrowing language.

For the case of inducing a Moroccan Darija-English translation lexicon via French we start with a French-English bilingual dictionary and take all the French pronunciations in IPA (International Phonetic Alphabet)[1] and convert them to how they would be rendered in Arabic script. For this we created a multiple-step transliteration process:

Step 1 Break pronunciation into syllables.

Step 2 Convert each IPA syllable to a string in modified Buckwalter transliteration[2], which supports a one-to-one mapping to Arabic script.

Step 3 Convert each syllable's string in modified Buckwalter transliteration to Arabic script.

Step 4 Merge the resulting Arabic script strings for each syllable to generate a candidate loanword string.

[1] https://en.wikipedia.org/wiki/International_Phonetic_Alphabet

[2] The modified version of Buckwalter transliteration, https://en.wikipedia.org/wiki/Buckwalter_transliteration, replaces special characters such as < and > with alphanumeric characters so that the transliterations are safe for use with other standards such as XML (Extensible Markup Language). For more information see (Habash, 2010).

For syllabification, for many word pronunciations the syllables are already marked in the IPA by the '.' character; if syllables are not already marked in the IPA, we run a simple syllabifier to complete step 1. For step 2, we asked a language expert to give us a sequence of rules to convert a syllable's pronunciation to modified Buckwalter transliteration. This is itself a multi-step process (see next paragraph for details). In step 3, we simply do the one-to-one conversion and obtain Arabic script for each syllable. In step 4, we merge the Arabic script for each syllable and get the generated candidate loanword string.

The multi-step process that takes place in step 2 of the process is:

Step 2.1 Make minor vowel adjustments in certain contexts, e.g., when 'a' is between two consonants it is changed to 'A'.

Step 2.2 Perform bulk of conversion by using table of mappings from IPA characters to modified Buckwalter characters such as 'a'→'a','k'→'k', 'y:'→'iy', etc. that were supplied by a language expert.

Step 2.3 Perform miscellaneous modifications to finalize the modified Buckwalter strings, e.g., if a syllable ends in 'a', then append an 'A' to that syllable.

The entire conversion process is illustrated in Figure 1 for the French word *raconteur*. At the top of the Figure is the IPA from the French dictionary entry with syllables marked. At the next level, step 1 (syllabification) has been completed. Step 2.1 doesn't apply to any of the syllables in this word since there are no minor vowel adjustments that are applicable for this word so at the next level each syllable is shown after step 2.2 has been completed. The next level shows the syllables after step 2.3 has been completed. The next level shows after step 3 has been completed and then at the end the strings are merged to form the candidate loanword.

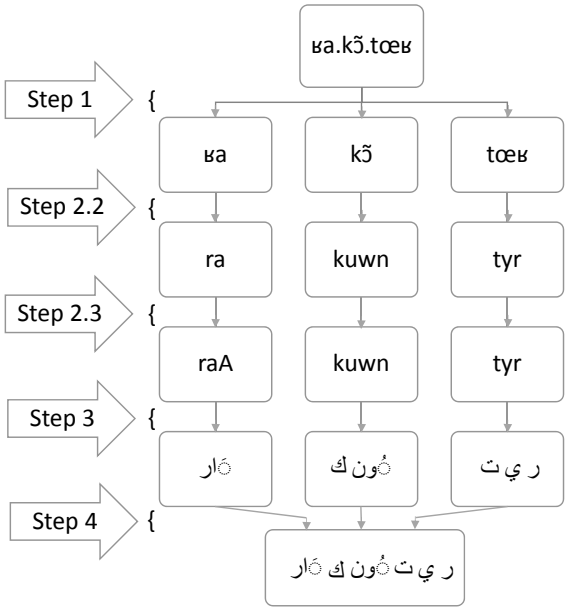

Figure 1: Example of French to Arabic Process for the French word *raconteur*. As discussed in the main text, step 2.1 doesn't apply to this example so it is omitted from the diagram to conserve space. Note that in the final step the word is in order of Unicode codepoints. Then application software that is capable of processing Arabic will render that as a proper Arabic string in right-to-left order with proper character joining adjustments as راكونتير

4 Experiments and Discussion

In our experiments we extracted a French-English bilingual dictionary using the freely available English Wiktionary dump 20131101 downloaded from http://dumps.wikimedia.org/enwiktionary. From this dump we extracted all the French words, their pronunciations, and their English definitions. Using the process described in section 3 to convert each of the French pronunciations into Arabic script yielded 8277 unique loanword candidate strings.

The data used for testing consists of a million lines of user comments crawled from the Moroccan news website http://www.hespress.com. The crawled user comments contain Moroccan Darija in heavily code-switched environments. While this makes for a challenging setting, it is a realistic representation of the types of environments in which historically unwritten languages are being written for the first time. The data we used is consistent with well-known code-switching among Arabic speakers, extending spoken discourse into formal writing (Bentahila and Davies, 1983; Redouane, 2005). The total number of tokens in our Hespress corpus is 18,781,041. We found that 1150 of our 8277 loanword candidates appear in our Hespress corpus. Moreover, more than a million (1169087) loanword candi-

Annotator	Arabic	Unknown	French	Total
A	907	88	190	1185
B	812	174	199	1185

Table 1: Number of word instances annotated.

date instances appear in the corpus. Recall that a match could be a true match that really is a French loanword or a false match that just happens to coincidentally have string equality with words in the borrowing language, but is not a French loanword. False matches are particularly likely to occur for very short words. Accordingly, we filter out candidates that are of length less than four characters. This leaves us with 838 candidates appearing in the corpus and 217616 candidate instances in the corpus. To get an idea of what percentage of our matches are true matches versus false matches, we conducted an annotation exercise with two native Moroccan Darija speakers who also knew at least intermediate French. We pulled a random sample[3] of 1185 candidate instances from our corpus and asked each annotator to mark each instance as either:

A if the instance is originally from Arabic,

F if the instance is originally from French, or

U if they were not sure.

The results are shown in Table 1. There are a substantial number of French loanwords that are found. Some examples of translations successfully induced by our method are:

omelette اومليت; and

bourgeoisie بورجوازي.

We hypothesize that our method can help improve machine translation (MT) of historically unwritten dialects with nearly zero resources. To test this hypothesis, we ran an MT experiment as follows.

First we selected a random set of sentences from the Hespress corpus that each contained at least one candidate instance and had an MSA/Moroccan Darija/English trilingual translator translate them into English. In total, 273 sentences were translated. This served as our test set.

[3]We removed 15 Arabic stopwords from our candidate list before pulling the random sample.

We trained a baseline MT system using all GALE MSA-English parallel corpora available from the Linguistic Data Consortium (LDC) from 2007 to 2013.[4]

We trained the system using Moses 3.0 with default parameters. This baseline system achieves BLEU score of 7.48 on our difficult test set of code-switched Moroccan Darija and MSA.

We trained a second system using the parallel corpora with our induced Moroccan Darija-English translation lexicon appended to the end of the training data. This time the BLEU score increased to 8.11, a gain of .63 BLEU points.

5 Conclusions

With the explosive growth of informal textual electronic communications such as social media, web comments, etc., many colloquial everyday languages that were historically unwritten are now being written for the first time often in heavily code-switched text with traditionally written languages. The new written versions of these languages pose significant challenges for multilingual processing technology due to Out-Of-Vocabulary (OOV) challenges. Yet it is relatively common that these historically unwritten languages borrow significant amounts of vocabulary from relatively well resourced written languages. We presented a method for translation lexicon induction via loanwords for alleviating the OOV challenges in these settings where the borrowing language has extremely limited amounts of resources available, in many cases not even substantial amounts of monolingual data that is typically exploited by previous cognates and loanword detection methods to induce translation lexicons. This paper demonstrates induction of a Moroccan Darija-English translation lexicon via bridging French loanwords using the method and in MT experiments, the addition of the induced Moroccan Darija-English lexicon increased system performance by .63 BLEU points.

Acknowledgments

We would like to thank Tim Buckwalter for his support and for providing us with the initial mapping of IPA syllables to their corresponding Arabic orthographies as well as the contextual adjustment rules that we used in our experiments.

[4]The LDC catalog numbers for the corpora we used are: LDC2008T09, LDC2007T24, LDC2008T02, LDC2009T09, LDC2009T03, LDC2012T14, LDC2012T06, LDC2012T17, LDC2012T18, LDC2013T01, and LDC2013T14.

References

Abdelali Bentahila and Eirlys E Davies. 1983. The syntax of Arabic-French code-switching. *Lingua* 59(4):301–330.

Michael Bloodgood and Benjamin Strauss. 2017. Using global constraints and reranking to improve cognates detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Vancouver, Canada. https://arxiv.org/abs/1704.07050.

Nizar Habash. 2010. *Introduction to Arabic Natural Language Processing*. Synthesis lectures on human language technologies. Morgan & Claypool Publishers. https://books.google.com/books?id=kRIHCnC74BoC.

Ann Irvine and Chris Callison-Burch. 2013. Supervised bilingual lexicon induction with multiple monolingual signals. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Atlanta, Georgia, pages 518–523. http://www.aclweb.org/anthology/N13-1056.

Grzegorz Kondrak, Daniel Marcu, and Kevin Knight. 2003. Cognates can improve statistical translation models. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume of the Proceedings of HLT-NAACL 2003–short Papers - Volume 2*. Association for Computational Linguistics, Stroudsburg, PA, USA, NAACL-Short '03, pages 46–48. http://www.aclweb.org/anthology/N/N03/N03-2016.pdf.

M. Paul Lewis, Gary F. Simons, and Charles D. Fennig. 2015. *Ethnologue: Languages of the world*, volume 18. SIL international, Dallas, TX.

Gideon S. Mann and David Yarowsky. 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*. Association for Computational Linguistics, Stroudsburg, PA, USA, NAACL '01, pages 1–8. http://www.aclweb.org/anthology/N/N01/N01-1020.pdf.

Benjamin S. Mericli and Michael Bloodgood. 2012. Annotating cognates and etymological origin in Turkic languages. In *Proceedings of the First Workshop on Language Resources and Technologies for Turkic Languages*. European Language Resources Association, Istanbul, Turkey, pages 47–51. http://arxiv.org/abs/1501.03191.

Rabia Redouane. 2005. Linguistic constraints on codeswitching and codemixing of bilingual Moroccan Arabic-French speakers in Canada. In *ISB4: Proceedings of the 4th International Symposium on Bilingualism*. pages 1921–1933.

Clinton Robinson and Karl Gadelii. 2003. Writing unwritten languages, a guide to the process. *Paris: UNESCO. Retrieved June 24, 2008* .

Charles Schafer and David Yarowsky. 2002. Inducing translation lexicons via diverse similarity measures and bridge languages. In *Proceedings of the 6th Conference on Natural language Learning - Volume 20*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 1–7. http://www.aclweb.org/anthology/W/W02/W02-2026.pdf.

Stephen Tratz, Douglas Briesch, Jamal Laoudi, and Clare Voss. 2013. Tweet conversation annotation tool with a focus on an arabic dialect, moroccan darija. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Association for Computational Linguistics, Sofia, Bulgaria, pages 135–139. http://www.aclweb.org/anthology/W13-2317.

Yulia Tsvetkov, Waleed Ammar, and Chris Dyer. 2015. Constraint-based models of lexical borrowing. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Denver, Colorado, pages 598–608. http://www.aclweb.org/anthology/N15-1062.

Yulia Tsvetkov and Chris Dyer. 2015. Lexicon stratification for translating out-of-vocabulary words. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Association for Computational Linguistics, Beijing, China, pages 125–131. http://www.aclweb.org/anthology/P15-2021.