Acquisition of Translation Lexicons for Historically Unwritten Languages via Bridging Loanwords

Michael Bloodgood[1] Benjamin Strauss[2]

[1]Department of Computer Science
The College of New Jersey

[2]Department of Computer Science and Engineering
The Ohio State University

Building and Using Comparable Corpora Workshop, August 3, 2017

Outline

- Introduction and Motivation
- Loanword Candidate Generation Method
- Experiments
- Conclusions and Future Work

Summary

- With the explosive growth of informal electronic communications such as social media, web comments, text messaging, etc., historically unwritten languages are being written for the first time.
- For these languages, there are extremely limited resources such as translation lexicons available.
- We present a method for inducing portions of translation lexicons through the use of expert knowledge for these settings and quantify its effectiveness in experiments attempting to induce a Moroccan Darija-English translation lexicon via French loanwords.

Motivation

- Translation lexicons are a core resource used for multilingual processing of languages.
- Manual creation of translation lexicons by lexicographers is time-consuming and expensive.
- There are more than seven thousand languages in the world, many of which are historically unwritten (Lewis et al., Ethnologue, 2015).
- Many historically unwritten languages are being written for the first time with the explosive growth of informal electronic communications.

Past work

- There has been a lot of work on automating translation lexicon induction, including (Bloodgood and Strauss, ACL, Vancouver, CA, 2017)
- The best methods for automatic translation lexicon induction involve using many sources of information such as word context information (Rapp, 1995, 1999), word frequency information, temporal information (Klementiev and Roth, 2006), word burstiness information (Church and Gale, 1995), and phonetic information.
- The methods for automatic translation lexicon induction have various data requirements such as bilingual seed dictionaries and monolingual text coming from the same time period for each of the languages.

Challenges

- For historically unwritten languages that are just being written for the first time, there are often extremely limited resources of any type available, not even large amounts of monolingual text.
- The written data that can be obtained often has non-standard spellings and code-switching.
- The code-switching is sometimes 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).

Potential Solution

- Many historically unwritten languages borrow parts of their lexicons from more highly resourced written languages.
- It is often possible to find a language informant that can provide guidance for how sounds would be rendered in a written script if words were to be written.
- Our proposed method makes use of these facts to acquire parts of a translation lexicon quickly.

Outline

- Introduction and Motivation
- Loanword Candidate Generation Method
- Experiments
- Conclusions and Future Work

Loanword Candidate Generation Method (high level summary)

- Take word pronunciations from the donor language and convert them to how they would be borrowed in the borrowing language if they were to be borrowed.
- These are our candidate loanwords.

Loanword Candidate Possibilities

- There are three possible cases for a given generated candidate loanword:

  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 the borrowing language.

Use Case: Moroccan Darija-English translation lexicon via French

- Our use case is 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) and convert them to how they would be rendered in Arabic script via a multiple step transliteration process.

Multiple-step Transliteration Process

Step 1 Break pronunciation into syllables.

Step 2 Convert each IPA syllable to a string in modified Buckwalter transliteration, which is a commonly used transliteration scheme that supports a one-to-one mapping to Arabic script.

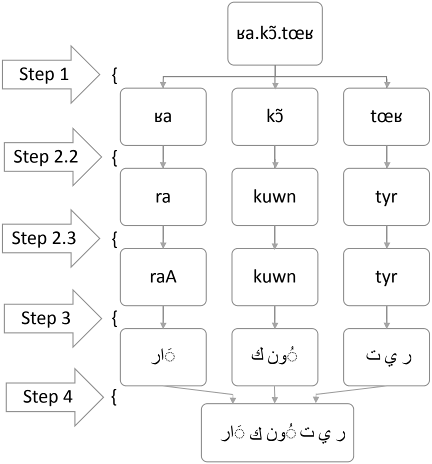
Step 3 Convert each syllable's string in modified Buckwalter transliteration to Arabic script.

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

Step 2

- 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.

Example of French to Arabic process for the French word *raconteur*



Outline

- Introduction and Motivation
- Loanword Candidate Generation Method
- Experiments
- Conclusions and Future Work

Experimental Data Sources

- We extracted a French-English bilingual dictionary using the freely available English Wiktionary dump 20131101 downloaded from `http://dumps.wikimedia.org/enwiktionary`.
- The data used for testing consists of a million lines of user comments crawled from the Moroccan news website `http://www.hespress.com`.

Initial Statistics of our Data

- Converting each of the French pronunciations from our dictionary into Arabic script yielded 8277 unique loanword candidates.
- 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.
- More than a million (1169087) loanword candidate instances appear in the corpus.

Filtering out short words

- False matches are particularly likely to occur for very short words.
- So 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.

Percentage of 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 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.

Annotation Results

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

Table: Number of word instances annotated.

Examples of Translations Found

omelette اوم ليت; and

bourgeoisie بورجوازي.

Machine Translation Experiment

- We selected a random set of sentences from the Hespress corpus that each contained at least one candidate instance.
- A Modern Standard Arabic/Moroccan Darija/English trilingual translator translated 273 of the sentences into English.
- These manually translated sentences served as our test set.
- We trained a baseline MT system using all GALE MSA-English parallel corpora available from the Linguistic Data Consortium from 2007 to 2013 using Moses 3.0 with default parameters.
- The baseline system achieves BLEU score of 7.48 on our difficult test set of code-switched Moroccan Darija and Modern Standard Arabic.
- We trained a second system with our induced translation lexicon appended to the end of the training data.
- The BLEU score increased to 8.11, a gain of 0.63 BLEU points.

Outline

- Introduction and Motivation
- Loanword Candidate Generation Method
- Experiments
- Conclusions and Future Work

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.
- The new written versions of these languages pose significant challenges for multilingual processing technology due to Out-Of-Vocabulary (OOV) challenges.
- Often 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.
- This paper demonstrates induction of a Moroccan Darija-English translation lexicon via bridging French loanwords using the approach.

Future Work

- Explore using the method for other languages.
- Examine whether adaptations can be made to increase the yield of the method.

Acknowledgment

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.

Questions

Questions