Key Features and Classifiers for English of Kazakhstan Compared to Other Asian Countries

Anonymous ACL submission

Abstract

English is the most spoken language in the entire world, with it being very popular for countries to have it as their second language, with nearly two thousand million speakers (Crystal, 2008). This implies that there are many different versions of it, and many idiosyncrasies within each version. This paper aims to demonstrate those of Kazakhstan, to gain a better understanding and hopefully come to an evaluation of a corpus of words that I will investigate, and hopefully this corpus will be representative of their national dialect. This will also be compared to other Asian countries in the hope of finding similarities, and differences.

1 Introduction

This paper aims to describe the process of gathering a corpus of English of Kazakhstan, the analysis of this corpus and using Machine Learning algorithms, the classification of the words to try and understand and justify its relevance and importance as a corpus. I will be using the analytical tools provided by the SketchEngine (Kilgarriff 2014; Kilgarriff et al. 2004) to try and extrapolate Kazakhstan's idiosyncrasies. I will be using N-grams and keyword extraction to try and get a better understanding of the national dialect. I will then be using the Waikato Environment for Knowledge Analysis, WEKA, to run machine learning classification algorithms on it to see which algorithms classify the corpus well and which ones do not, and from this derive how representative the corpus is of national dialect. I will also be comparing it to a corpus of English from Thailand, China, Nepal and India to see if there are country specific features or if there are many similarities within the analysis of the corpora. This investigation will follow the CRISP-DM format for clarity and simplicity, to ensure that the paper is fully understandable.

2 Literary Review

I wanted to read papers on this topic to try and understand the scope of the project better, and to see how researchers go about the process of using SketchEngine and WEKA. I read that extracting multi-word expressions carry a lot more meaning (S. Alrehaili and E. Atwell) than single words, so I believed this would give a good understanding of the dialect of Kazakhstan. I also saw that SOM was an algorithm that performs well in the classification of text (Tarmon, et al., 2020) so I knew that this was a potential algorithm for good classification results.

3 Business Understanding

The purpose of this investigation is to compare and contrast features of dialects of countries from similar geographical locations, and to try and classify samples of text from our country's dialect.

3.1 Objectives

Gathering a corpus of English of Kazakhstan can provide a unique perspective on the description of the language paraphrasing (Atwell, 1986). Once a corpus has been gathered, I will aim to find expressions and words that are specific to the dialect of Kazakhstan, and to see how well the text gathered from WebBootCat can be classified using Machine learning algorithms.

3.2 Success Criteria

Each country will be compared using classification algorithms that WEKA provides, such as J48, SOM, Naïve-Bayes. The algorithms will return which countries were able to classify the best, and which ones did not, by providing a classification percentage and confusion matrix. I will also be able to use SketchEngine's tools to try and find N-grams and keywords, and any distinctive patterns and expressions. If there are, this will be regarded as a successful outcome.

4 Data Understanding

Data is going to be scraped from the web, using the specified top-level domain. The data will require keywords for information retrieval, and these keywords must be selected in such fashion to ascertain a corpus that is indicative of a dialect. I have chosen words that are not often used (Brierley et al., 2013) as to try and gather as best a corpus as possible

4.1 Exploration

I tried many different words, but they all produced a corpus that was far too large to work with, so I had to be more specific. I chose the topic of education to centre my words around, to try and get a smaller corpus size. Initially I gathered a corpus of around 100,000 words but upon review of the sites and N-grams some of them were do with quizzes and statistics. I had to remove these due to the fact they are not relevant to linguistic analysis. The clearly erroneous data was now removed. However, as the corpus is small it may not be a big enough pool to conclusively find multi-word expressions (Pickard, 2020).

5 Data Preparation

Each corpus file contained a lot of unnecessary information, such as bullet points, html tags, and extra erroneous spaces. There were also words that were not in the Latin alphabet, but in Cyrillic or Chinese characters. These had to be removed as it is important to have a data set that was text and only text, as this would skew any analysis or interpretation of this data. I was then going to need to change it into a format which WEKA would recognise.

5.1 Cleaning the corpus

I created a python script that allowed me to remove almost all of the problematic data. I used regex to remove any HTML tags, and then used python's in-built string methods to replace any characters that were not fit for the text, including extra quotation marks that caused issues with the arff file format.

```
    entry = re.sub("<.*?>", "", entry)
    entry = entry.strip("\n")

    words = entry.split()
    numberWords += len(words)

    if numberWords < 4:
        entry = entry.replace(entry,'')
        fEdited.write(entry)
    else:
        fEdited.write('"' + entry + '",KZ \n')
```

5.2 Making the combined corpus

In order to compare and contrast different countries in WEKA, I needed to create a file for WEKA that contained all the data from every corpus, and have that data marked for each country. I wrote a different python script to do this. After cleaning up each corpus, I ran the script and it created one big corpus that was ready for classification and analysis.

6 Modelling

To achieve results that are accurate and representative of the data, I wanted to be very methodical and thorough in my approach to classification of the data. I chose 5 classifiers, Naïve-Bayes, Naïve-Bayes Multinomial Updateable, SOM, ZeroR and J48, all provided by WEKA. I ran both percentage split at 66% and cross-validation with ten folds, as I wanted to see how they compared. In SketchEngine I would be analysing the N-grams, and key words to see if I could get a better understanding of the grammatical expression and sentence structure of Kazakhstan's dialect, and hopefully find some idiosyncrasies. I will also be using WEKA to compare the classification of the different

countries. This was to see which ones classified better than others and try and draw some conclusions from that.

6.1 SketchEngine

Running the Keywords tool on Sketch Engine, which compares it to the English Web 2018 corpus (enTenTen18), I was able to see keywords and expressions that are common in my corpus. I found that, in single words, unusually common words were “preactivated”, “extra-financial” and “processes-level”. In terms of N-grams, I found “lifetime giving”, “poverty alleviation” and “engendering trust”. Upon further investigation of the single words, to try and understand their collocational behaviour, I found that “preactivated” is used in terms of digital keys, whereas “extra-financial” and “processes-level” were used as nouns. For the N-grams, “lifetime giving” was used extensively, with many variations including “entire lifetime”, and “total lifetime”.

6.2 Classifiers

I started off with using the StringToWordVector filter to be able to classify my corpus better. I ran all ZeroR, J48, Naïve-Bayes, NaïveBayesMultionomialUpdateable and SOM, with both percentage split and cross-validation options.

6.3 Classifier Outputs and Results

Classifier	Correct Classified Instances (% to 2 d.p.)	
	Percentage Split (66%)	Cross-Validation (10 Folds)
J48	68.30	65.44
Naïve-Bayes	54.64	53.85
SOM	77.49	79.74
NaiveBayesMultionomialUpdateable	80.48	81.25
ZeroR	36.11	36.11

Figure 1: Table of Results.

I achieved this table of results. I was able to classify the best and the worst with NaiveBayesMultionomialUpdateable and ZeroR respectively.

```
   a    b    c    d    e   <-- classified as
4238    0    0    0    0 |    a = TH
1114    0    0    0    0 |    b = CN
1064    0    0    0    0 |    c = IN
2214    0    0    0    0 |    d = KZ
3105    0    0    0    0 |    e = NP
```

Figure 2: ZeroR – Cross Validation.

ZeroR – Percentage Split

```
=== Confusion Matrix ===

   a    b    c    d    e   <-- classified as
1462    0    0    0    0 |    a = TH
 365    0    0    0    0 |    b = CN
 384    0    0    0    0 |    c = IN
 746    0    0    0    0 |    d = KZ
1033    0    0    0    0 |    e = NP
```

Figure 3: ZeroR – Percentage Split.

```
=== Confusion Matrix ===

   a    b    c    d    e   <-- classified as
3712   65   75  201  185 |    a = TH
  43  939   44   35   53 |    b = CN
  48   60  861   27   68 |    c = IN
 228   61   66 1720  139 |    d = KZ
 378   97  104  223 2303 |    e = NP
```

Figure 4: NaiveBayesMultionomialUpdateable – Cross Validation.

```
=== Confusion Matrix ===

   a    b    c    d    e   <-- classified as
1253   26   25   81   77 |    a = TH
  14  305   17   12   17 |    b = CN
  16   19  310    4   35 |    c = IN
  77   26   19  578   46 |    d = KZ
 139   30   25   78  761 |    e = NP
```

Figure 5: NaiveBayesMultionomialUpdateable – Percentage Split

Bar charts of the classifiers.

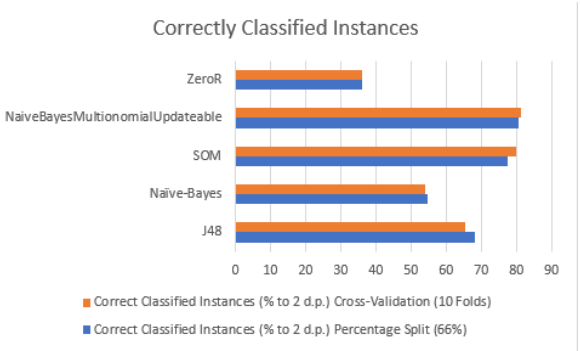

Figure 6: Bar Chart of the Results, Separated by Classifier.

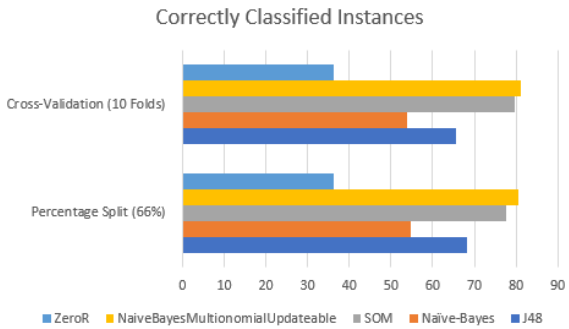

Figure 6: Bar Chart of the Results, Separated by test option.

Percentage Split Accuracy Mean – 63.41%
Cross-Validation Accuracy Mean – 63.28

7 Evaluation

7.1 Classifier Interpretation

From my classifier outputs, I was able to achieve a classification of 81.25% using NaiveBayesMultionomialUpdateable algorithm, and with cross-validation with 10 folds, with SOM coming closely behind it. ZeroR was clearly the worst classifier with only a 36.11% classification. As ZeroR is the baseline for the classifiers, we know this is why it performed the worst, as it has no predictability power. It also produced no comparison between the countries. I achieved the best results with NaiveBayesMultionomialUpdateable, and SOM closely behind it. NaiveBayesMultionomialUpdateable is an incremental version of NaiveBayesMultionomial, so the prediction results will get better the further the algorithm runs and hence it's success in classification. SOM is similar in principle and as it uses a heuristic approach so it too will learn as the algorithm goes on. This implies that the data set requires the algorithm to learn from itself, and that it may not itself be of a good standard, represented by the other algorithm's classification success results. Looking at the confusion matrix, India and Kazakhstan classified the least with other countries, which implies that they are the most unique data sets, with Thailand being the least unique. On the algorithms that performed the worst, ZeroR and Naïve-Bayes, percentage split was the test option that returned the best result, and the algorithms that performed the best, cross-validation was the best test option. I believe this means that when the classification leans to doing well it will continue to, and if it leans to doing badly it will continue to also.

7.2 SketchEngine Interpretation

I believe that my corpus did not produce anything that could be called an idiosyncrasy of Kazakhstan's dialect. It showed me that often they will put two words which they know have the correct meaning such as "lifetime giving", but an unusual way of saying it. This is not clear evidence of a unique dialect as the other countries I compared it to, had roughly similar statements which were too unusual. I also believe that a 50,000-word corpus per country is too small to produce an accurate representation of an entire dialect, and thus I struggled to find conclusively unique, idiosyncratic multi-word expressions. I also noticed that most of the data scraped was from official websites and educational websites, that have branches for each country, and therefore may not even be representative of Kazakhstan itself.

8 Conclusion

The classification results I believe were good for my corpus, as I think my corpus may have contained a lot of official websites which often used well-structured English, and this unfortunately I did not find any idiosyncrasies within my corpus. It classified well, with the best classifier with the best option producing a above 80% classification of instances, with Kazakhstan's being the second-most unique. In retrospect, longer time spent on achieving a better corpus would have produced better results, and maybe then I would have been able to find a corpus that classified well, was representative of their dialect, and produced some idiosyncrasies of their language.

References

Crystal, D. 2008 "Two thousand million?" English Today. Cambridge University Press.

S. Alrehaili and E. Atwell. 2017. Extraction of Multi-Word Terms and Complex Terms from the Classical Arabic Text of the Quran. *International Journal on Islamic*

Applications in Computer Science And Technology. 5(3), pp. 15-27.

Kilgarriff, A. 2014. Sketch engine [Computer Software]. Retrieved April 26, 2021, from http://www.sketchengine.co.uk/.

E, Atwell. 1986. A parsing expert system which learns from corpus analysis. *Proceedings of ICAME'1986 International Computer Archive of Modern English conference*, pp. 227-232

C. Brierley, E. Atwell, C. Rowland, and J. Anderson. 2013. Semantic pathways: a novel visualisation of varieties of English. *ICAME Journal of the International Computer Archive of Modern English.* 37, pp. 5-36.

T. Pickard. 2020. Comparing word2vec and GloVe for Automatic Measurement of MWE Compositionality. In *Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons.* pp. 95-100.