Instructions for ACL-2014 Proceedings

Mihael Simoni
Kognitive Science
University of Tbingen

Nikolas Zeitler
Computer Engineering
University of Tbingen

Abstract

LALALA abstract

1 Introduction

In the European-Union learn 94.5% of the population english in school. In some areas like the Netherlands or the Czech Republic eveb 100%(0). It is due to this clear that english is used as an bridge-language to communicate between europeans from different countries.

Through new technologies and the fall away of border is it easier then ever to communicate with people from different countries. And of course we use english more and more on a daily basis. In the german language are words like sexy, laptop, fast food, wellness and many more widely spread[CITE FEHLT]. Furthermore could advertising not live without english anymore (e.G. Douglas - Your partner in beauty). The german language goes even a step further and invents its own english words (e.G Handy is german for mobile-phone).

It is therefore interesting to look at code-switching between english and other languages. In this paper we will look at tweets containing english and spanish words. Our goal is it to find out the language for each word.

In this paper we perform a language identification using machine learning techniques. Those techniques are able to indeficate the language of a single word without looking at the sentence. After the machine learning algorithm has indetificated the languages of each word, we try to improve the result using statistical facts of code-switching. We have with example observed that code-switching occurs very rarely in between two other words of the same language.

VERGLEICH VERSCHIEDENER ANSTZE? LINEARE REGRESSION / NEURONALE NETZE

The data we are using are provided by *First Workshop on Computational Approaches to Code Switching*(0). We have 11.400 Spanish-English tweets. Some of them are deleted or invalid thus we have XXX tweets and XXX different words. We know the language of each word in every tweet.

To evaluate our methods the workshop(0) provided us with addional test- and trial-tweets. SECTION X DOES LALALALA SECTION Y DOES LALALALA

2 Data

The data provided by (0) contains english and spanish tweets. For each tweet are provided with how to split the words and which language this word is. Tweets contain a lot of special characters. Thus the data not only has labels for english *lang1* and spanish *lang2*. But also *other* for emoticons, nicknames or gibberish. With example ":)", "@Ody12", or "Zaaaas" are labeled as other. The lable *mixed* describes if a word contains both, spanish and english words, "ClutterDesordenLook" is such a word. The next label is *ne*, ne is a name entity. Those are proper names refering to people, places, organizations, locations or titles. The difficulty for those is that they could span above multiple words "West Coast" for example.

The data provides information of how to split the tweet into words. This is important in cases where punctuation marks or emoticons are directly at a word without whitespace in between with example "HEY!:)".

Lets say the data gives us the tweet: "@snapchateame No! Jason you look bueno!:). Next thing provides the given data the information to split the tweet into words to look like this: "snapchateame, No, Jason, you, look, bueno, :)". And finaly the data tells us which label every word gets "mixed, ambiguous, ne, lang1, lang1, lang1, lang2, other”.

3 Approaches

We want to find the language for every word inside a tweet. For this we use a two step approach. The first step is to train and use a machine learning algorithm. And the next step is to check its results and improve it using our knowledge about code-switching and statistical approximations.
Goal of the machine learning algorithm is it to look at a single word and decide which language it is. Thus we use uni- and bigrams to train the algorithm. The unigrams are every character used in the tweets, and the bigrams is a combination of every charachter with every character. After we obtained the uni- and bigrams we looked up every distinct word in the provided tweets. Afterwards we wrote in a table which uni- or bigrams are occurring in each word. Additional we provided the table with the language from which the word is.

4 Evaluation

5 Summary

References

Conference on Empirical Methods in Natural Language Processing ,*First Workshop on Computational Approaches to Code Switching*, *emnlp*2014*.org/workshops/CodeSwitch/call.html*, Doha, Qatar, October 2014.

eurostat - Your key to European statistics, *Foreign languages learnt per pupil in upper secondary education, 2007 and 2012*, June 2015.