Homework Four: Natural Language Processing

Quin'darius Lyles-Woods

November 19, 2021

1 Lexical Translation Model using IBM Model 1

Given a corpus of two English sentences with their translation in Latin below.

English this course

Latin hoc utique

English Text processing

Latin Textus processus

Build your lexical translation model (IBM Model 1) and show each training (expectation maximization) step's computation. Assume that the model only trains for two iterations.

Uniform Probablitiy

$$Probablity(this|hoc) = \frac{1}{2} \qquad Probablity(Text|Textus) = \frac{1}{2}$$
$$Probablity(course|utique) = \frac{1}{2} \qquad Probablity(processing|processus) = \frac{1}{2}$$

$$Probablity(this|hoc) * Probablity(course|utique) = \frac{1}{4}$$
$$Probablity(Text|Textus) * Probablity(processing|processus) = \frac{1}{4}$$

$$Probablity(A|E,F) = \frac{1}{2}$$
$$Probablity(A|E,F) = \frac{1}{2}$$

2 Translation Model using BLUE Score

Given candidate translation sentences and their references below. Evaluate your translation model performance using BLUE score where n-gram order, $N = 3$.

2.1 Answer

Given the N Gram count of 3 it makes the model not pickup the phrases well. The model doesn't match any of the N grams in the second candidate sentence.

Blue Score, Replacing ones with zeros by transforming all values by one.

$$Sentence\ One = \left\{ \frac{8}{8} * \frac{5}{7} * \frac{3}{6} \right\}^{\frac{1}{3}} = .7095$$

$$Sentence\ Two = \left\{ \frac{6}{6} * \frac{3}{5} * \frac{1}{4} \right\}^{\frac{1}{3}} = .5313$$

$$Average = \left\{ \frac{.7095 + .5315}{2} \right\} = 62.04\%$$

Candidate One the cat the cat on the mat

Unigram
R1 R2 - the
R1 R2 - cat
R1 R2 - the
R1 R2 - cat
R1 R2 - on
R1 R2 - the
R1 R2 - mat
Bigram
R1 R2 - the cat
R1 R2 - cat the
R1 R2 - the cat
R1 R2 - cat on
R1 R2 - on the
R1 R2 - the mat
Trigram
R1 R2 - the cat the
R1 R2 - cat the cat
R1 R2 - the cat on
R1 R2 - cat on the
R1 R2 - on the mat

Candidate Two the cat on mat mat

Unigram
R1 R2 - the
R1 R2 - cat
R1 R2 - on
R1 R2 - mat
R1 R2 - mat
Bigram
R1 R2 - the cat
R1 R2 - cat on
R1 R2 - on mat
R1 R2 - mat mat
Trigram
R1 R2 - the cat on
R1 R2 - cat on mat
R1 R2 - on mat mat

Reference One the cat is on the mat

Unigram
C1 C2 - the
C1 C2 - cat
C1 C2 - is
C1 C2 - on
C1 C2 - the
C1 C2 - mat
Bigram
C1 C2 - the cat
C1 C2 - cat is
C1 C2 - is on
C1 C2 - on the
C1 C2 - the mat
Trigram
C1 C2 - the cat is
C1 C2 - cat is on
C1 C2 - is on the
C1 C2 - on the mat

Reference Two there is a cat on the mat

Unigram
– there
C1 C2 - is
C1 C2 - a
C1 C2 - cat
C1 C2 - on
C1 C2 - the
C1 C2 - mat
Bigram
C1 C2 - there is
C1 C2 - is a
C1 C2 - a cat
C1 C2 - cat on
C1 C2 - on the
C1 C2 - the mat
Trigram
C1 C2 - there is a
C1 C2 - is a cat
C1 C2 - a cat on
C1 C2 - cat on the
C1 C2 - on the mat

3 Passage Retrieval

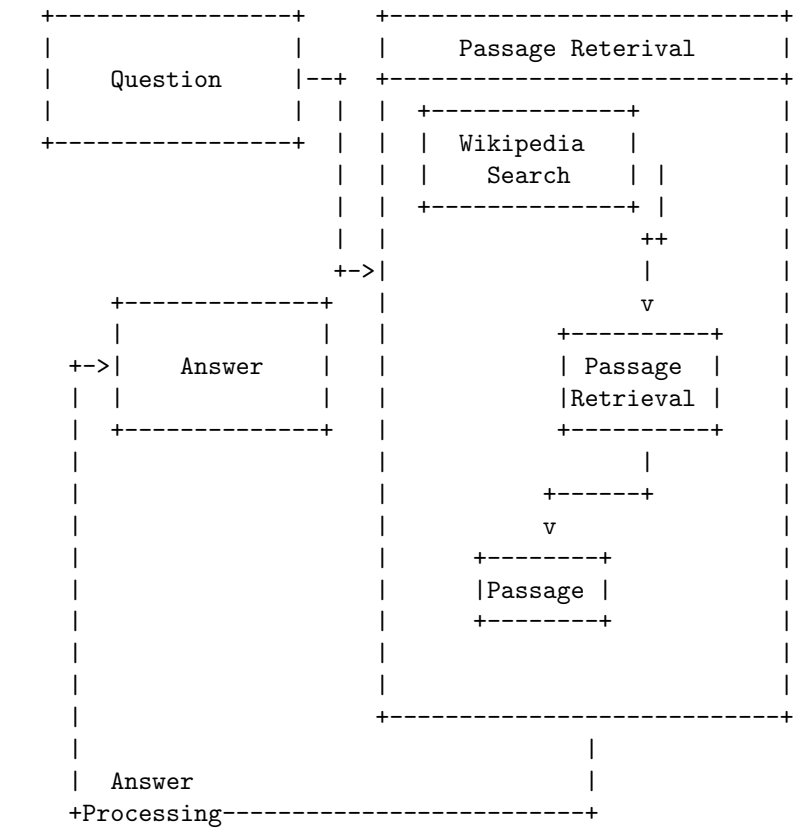
Design a question answering system for finding definition of all the machine learning keywords from Wikipedia pages.

Example Query Text: Support Vector Machine

Your system should return the corresponding wiki page for this query. Please show overall diagram of your system and explain each individual component.

Answer

There is a program called docgrab that I have created that will reach out a reformat the question into a query for wiki. Once that is done it will search and return the appropriate article. Then the program output is piped into another program that will get the HTML document and return only the text describing the search query.




Code
quindarius@Natural-Language-Processing/homework_four/Code
./docgrab "What is Support vector machine?" | pup 'div.mw-parser-output p:nth-child(4) text{}'
Code
that analyze data for
classification
and
regression analysis
. Developed at
AT&T Bell Laboratories
by
Vladimir Vapnik
with colleagues (Boser et al., 1992, Guyon et al., 1993, Vapnik et al., 1997
[
citation needed
]
) SVMs are one of the most robust prediction methods, being based on statistical
learning frameworks or
VC theory
proposed by Vapnik (1982, 1995) and Chervonenkis (1974). Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-
probabilistic
binary
linear classifier

4 Entity Linking

What is entity linking? How the overall entity linking system works? How would you find matching between mention and entity in the linking system?

Answers

Entity Linking

Entity Linking is the task of assigning a unique identity to entities mentioned in text.

Entity Linking System

Linking parts of a passage together requires a dataset that contains lots of known entities. You can parse through a document finding the comparing each of the unique tokens to the dataset that you have on and mark the ones that match as entities is distinct categories such as people, places, events.

Matching in a linking system

A very simple but robust version of this is using Wikipedia's articles as your data set. Simply if the article is not there its probably not going to be an entity.

5 Machine Translation Issues

What is the machine translation issues? Please provide explanations and examples for each of the cases.

Answer

The problem with machine translation are lexical translation, and reordering.

Lexical Translation

Lexical Divergence[1] , cause words that could mean different things to produce erroneous outcomes within machine translation. The parts of speech that exposes this most are going to be the Homonym[2] and Polysemous[3].

Reordering

Local reordering deals with the order of operations of the part of speech. While long distance reordering deals with the complete syntactic parts of speech.

[1] Ambiguities with the language.

[2] each of two or more words having the same spelling or pronunciation but different meanings and origins.

[3] the coexistence of many possible meanings for a word or phrase.