# Objective

To disambiguate word pairs using a Naive-Bayesian technique and answer to some questions.

# Installation

**Programming language:** Python

**Source code location:** /home/sbillah/nlp2/

**Corpus Selection:** I select "**senseval"** corpus, which is specially designed for WSD. Here is the download link: **http://www.senseval.org/.**

# Design (Naive Bayesian Approach)

## Algorithm:

I use two in-memory Hash-tables to store conditional probabilities of contexts associated with the pseudowords.

*// preprocess corpus files.*

*for each sense-file f:*

1. *replace the* ***word*** *with pseudoword and remember its sense in <tag> region.*
2. *extract the context words in both side of the pseudoword, and*

*store the contexts in two files: training, and testing by 8:2 ratio.*

*//training*

*For each training file f:*

*for each line in f:*

1. *update C(context-words), C(word), & sense sk in respected hash-tables.*

*from the counts, compute P(ci|sk), p(sk)and store in hash-tables.*

*//testing*

*for each testing file f:*

*for each line in f:*

1. *apply Laplace smoothing on conditional probabilities.*
2. *compute argmax score(sk) using the formula in the book.*

*3. compare the predicted value with actual value.*

*return accuracy in percentage.*

**Time Complexity:**

* **Preprocessing phase:** O(# of lines containing word1) + O(#of lines containing word2)
* **Training phase:** 2\*O(2\*context\_size \* #lines in training file)
* **Testing phase:** O(2\*context\_size \* #lines in testing file)
* **Overall:** O(2\*context\_size\*(#of word1+ #of word2)

**Space Complexity:**

* **Overall:** O(2 \* context\_size \* (#of word1 + #of word2) )

## Corpus Description:

The **Senseval** WSD corpus has total 35 sense-tagged words. Each word has more than 5 senses. But due to the simplified requirement of our homework, I ignore all those senses. Therefore, for a word pair, I consider only two senses (0,1). Here are my selected word-pairs and their individual occurrence in the corpus.

|  |  |  |
| --- | --- | --- |
| **Pair** | **Words** | **Word Counts** |
| 1 | amaze | 319 |
| behaviour | 1003 |
| 2 | sack | 296 |
| sanciton | 101 |
| 3 | knee | 477 |
| onion | 29 |
| 4 | accident | 1303 |
| wooden | 370 |

Below is a snapshot of some lines from "***accident.cor****"* file (context for word, 'accident'):

*800001*

*Late on Thursday night it was travelling at about three metres a second in wind blowing at 20 to 25 knots when an empty car fell off just as it reached the top.*

*The* ***<tag "532675">accident</>*** *appeared to have little effect on the Christmas party, except to lengthen it considerably.*

*800002*

*An image of earnest Greenery is almost tangible.*

*Eighteen years ago she lost one of her six children in an* ***<tag "532675">accident</>*** *on Stratford Road, a tragedy which has become a pawn in the pitiless point-scoring of small-town vindictiveness.*

800003

It's a sentiment I recommend to you all.

The **<tag "532675">accident</>** occurred on the Saturday of the annual Popular Flying Association (PFA) rally at Cranfield.

...

## Context Selection:

I varied context length from 1 to 19 (on both side) as shown in the figure below:

Context size

Context size

Different level of accuracy is obtained under different context size. The results are given in the next chapter.

## Laplace Smoothing:

During testing phase, some context-words are not seen before in training phase. Instead of assigning zero probability for them, I use Laplace Smoothing. The Laplace Smoothing is given below:

* If P(ci|s1) > 0:
  + *P(ci|s1)' = (P(ci|s1)\*10000+1)/( 10000 + context\_size)*
* Else:
  + *P(ci|s1)' = 1.0/( 10000 + context\_size)*

# Experimental Result

Here, I provide the experimental results for different context size and word-pairs. After the table, I also present these data graphically for better understanding.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Context Size | Pair 1 accuracy  {accident, wooden} | Pair 2 accuracy  {knee, onion} | Pair 3 accuracy  {sack, sanction} | Pair 4 accuracy  {amaze, behaviour} |
| 1 | 88.09524 | 95.65217 | 85.48387 | 88.04781 |
| 2 | 89.68254 | 100 | 85.48387 | 88.04781 |
| 3 | 89.28571 | 100 | 77.41935 | 83.66534 |
| 4 | 90.47619 | 100 | 77.41935 | 82.47012 |
| 5 | 90.07937 | 100 | 79.03226 | 82.07171 |
| 6 | 90.87302 | 100 | 79.03226 | 81.67331 |
| 7 | 88.88889 | 100 | 85.48387 | 83.26693 |
| 8 | 86.50794 | 100 | 87.09677 | 82.86853 |
| 9 | 86.90476 | 100 | 82.25806 | 81.2749 |
| 10 | 88.88889 | 100 | 82.25806 | 80.47809 |
| 11 | 87.69841 | 97.82609 | 82.25806 | 80.07968 |
| 12 | 87.69841 | 97.82609 | 80.64516 | 80.87649 |
| 13 | 86.90476 | 97.82609 | 82.25806 | 79.68127 |
| 14 | 84.92063 | 97.82609 | 82.25806 | 80.87649 |
| 15 | 85.31746 | 97.82609 | 80.64516 | 78.88446 |
| 16 | 84.52381 | 97.82609 | 80.64516 | 78.88446 |
| 17 | 83.73016 | 97.82609 | 79.03226 | 77.68924 |
| 18 | 82.14286 | 95.65217 | 79.03226 | 77.29084 |
| 19 | 81.74603 | 97.82609 | 75.80645 | 77.68924 |

|  |
| --- |
|  |
|  |
|  |
|  |

# Discussion:

1. Accuracy always decrease with the increase of context size.
2. For different word-pairs, maximum accuracy is obtained in different context size.
3. The overall performance of Naive Bayesian disambiguation is around 90%.

# Q&A

2.9 Relative frequency:

**File:** DavidBowie.html

**Entropy:** 3.52728218653

|  |  |
| --- | --- |
| 1 | 2370 |
| 0 | 1909 |
| 3 | 885 |
| 2 | 1502 |
| 5 | 784 |
| 4 | 750 |
| 7 | 881 |
| 6 | 721 |
| 9 | 1358 |
| 8 | 917 |
| a | 19075 |
| c | 8595 |
| b | 4576 |
| e | 23268 |
| d | 8301 |
| g | 3754 |
| f | 5294 |
| i | 20024 |
| h | 7057 |
| k | 3306 |
| j | 423 |
| m | 4891 |
| l | 11101 |
| o | 11614 |
| n | 13074 |
| q | 144 |
| p | 7083 |
| s | 14664 |
| r | 12658 |
| u | 4556 |
| t | 16047 |
| w | 5067 |
| v | 2178 |
| y | 2469 |
| x | 1038 |
| z | 232 |

**File:** Genghis Khan - Wikipedia, the free encyclopedia.html

**Entropy:** 3.55870437955

|  |  |
| --- | --- |
| 1 | 1844 |
| 0 | 1870 |
| 3 | 873 |
| 2 | 1562 |
| 5 | 686 |
| 4 | 566 |
| 7 | 505 |
| 6 | 610 |
| 9 | 624 |
| 8 | 594 |
| a | 20479 |
| c | 7362 |
| b | 3328 |
| e | 21399 |
| d | 7111 |
| g | 6384 |
| f | 4749 |
| i | 20236 |
| h | 9353 |
| k | 4675 |
| j | 848 |
| m | 5568 |
| l | 11457 |
| o | 11204 |
| n | 14745 |
| q | 263 |
| p | 6122 |
| s | 13442 |
| r | 11698 |
| u | 4415 |
| t | 15763 |
| w | 4265 |
| v | 1788 |
| y | 2359 |
| x | 1218 |
| z | 392 |

**File:** Steve Jobs - Wikipedia, the free encyclopedia.html

**Entropy:** 3.61251296863

|  |  |
| --- | --- |
| 1 | 7046 |
| 0 | 7798 |
| 3 | 1564 |
| 2 | 6815 |
| 5 | 1808 |
| 4 | 1605 |
| 7 | 1667 |
| 6 | 1966 |
| 9 | 2120 |
| 8 | 1811 |
| a | 36239 |
| c | 18155 |
| b | 7692 |
| e | 47051 |
| d | 11058 |
| g | 6116 |
| f | 10656 |
| i | 31138 |
| h | 14120 |
| k | 4429 |
| j | 2588 |
| m | 9194 |
| l | 22943 |
| o | 23755 |
| n | 26050 |
| q | 179 |
| p | 16747 |
| s | 31064 |
| r | 23907 |
| u | 6540 |
| t | 32956 |
| w | 9403 |
| v | 5120 |
| y | 4506 |
| x | 3221 |
| z | 504 |

**File:** Winston Churchill - Wikipedia, the free encyclopedia.html

**Entropy:** 3.54645118781

|  |  |
| --- | --- |
| 1 | 6843 |
| 0 | 5721 |
| 3 | 1761 |
| 2 | 3863 |
| 5 | 2560 |
| 4 | 1703 |
| 7 | 1273 |
| 6 | 1797 |
| 9 | 2795 |
| 8 | 1769 |
| a | 45163 |
| c | 21142 |
| b | 9324 |
| e | 56370 |
| d | 19968 |
| g | 9962 |
| f | 13635 |
| i | 49316 |
| H | 21543 |
| K | 8332 |
| J | 1441 |
| M | 12329 |
| L | 34877 |
| O | 29943 |
| N | 34098 |
| Q | 445 |
| P | 16278 |
| s | 34129 |
| r | 35857 |
| u | 10682 |
| t | 43826 |
| w | 12176 |
| v | 5852 |
| y | 7013 |
| x | 3895 |
| z | 654 |

**Finally, the total Corpus frequency:**

|  |  |
| --- | --- |
| 1 | 18103 |
| 0 | 17298 |
| 3 | 5083 |
| 2 | 13742 |
| 5 | 5838 |
| 4 | 4624 |
| 7 | 4326 |
| 6 | 5094 |
| 9 | 6897 |
| 8 | 5091 |
| a | 120956 |
| c | 55254 |
| b | 24920 |
| e | 148088 |
| d | 46438 |
| g | 26216 |
| f | 34334 |
| i | 120714 |
| H | 52073 |
| K | 20742 |
| J | 5300 |
| M | 31982 |
| L | 80378 |
| O | 76516 |
| N | 87967 |
| Q | 1031 |
| P | 46230 |
| s | 93299 |
| r | 84120 |
| u | 26193 |
| t | 108592 |
| w | 30911 |
| v | 14938 |
| y | 16347 |
| x | 9372 |
| z | 1782 |

**2.10 KL Divergence**

|  |  |
| --- | --- |
| KL-divergence <1,2> | 0.003688 |
| KL-divergence <2,1> | 0.003688 |
| **KL-divergence <1,3>** | **0.03306** |
| **KL-divergence <3,1>** | **0.03306** |
| KL-divergence <2,3> | 0.028468 |
| KL-divergence <3,2> | 0.028468 |

So, the corpus 1(english1) and corpus 3 (french1) have the highest score. In fact, these two corpus are same and translation of each other, which justifies the result.