STRUCTURE PREDICTION IN RECIPE INGREDIENT

SANKET DESHMUKH, SRINIVAAS GANESAN

Main motivation behind the project is to extract the structure of a recipe ingredient. For this we implement a HMM and Structured Perceptron and compare the results

DATASET

- We are using NewYork Times cooking recipe dataset which has about 179,000 entries
- We are splitting it into 70% training and 30% test data
- Each entry is an ingredient for eg. '2 to 3 cups thinly sliced onion'
- The task is to assign tags for them. We have 5 tags. *Qty_start* : measure of minimal units needed, *Qty_end* : measure of maximal units needed, *Name* : the main ingredient, *Unit* : the unit of measure, *Comment* : any comments on the ingredient

TAG ASSIGNMENT
2 – Qty_start
3 – Qty_end
Cups – Unit
thinly– Comment
sliced – Comment
Onion - Name

HIDDEN MARKOV MODEL

- We calculate Emission and transition probabilities with the dataset. Stop words are not considered
- Using Viterbi algorithm we predict the best tag sequence for a sentence
- As Qty_s and Qty_e are numeric we parse them manually and are not tagged by the HMM
- In case of unseen words, Transition probabilities alone are used to predict the state
- HMM gives pretty good word level accuracy and a sentence level accuracy of **58%** in test data

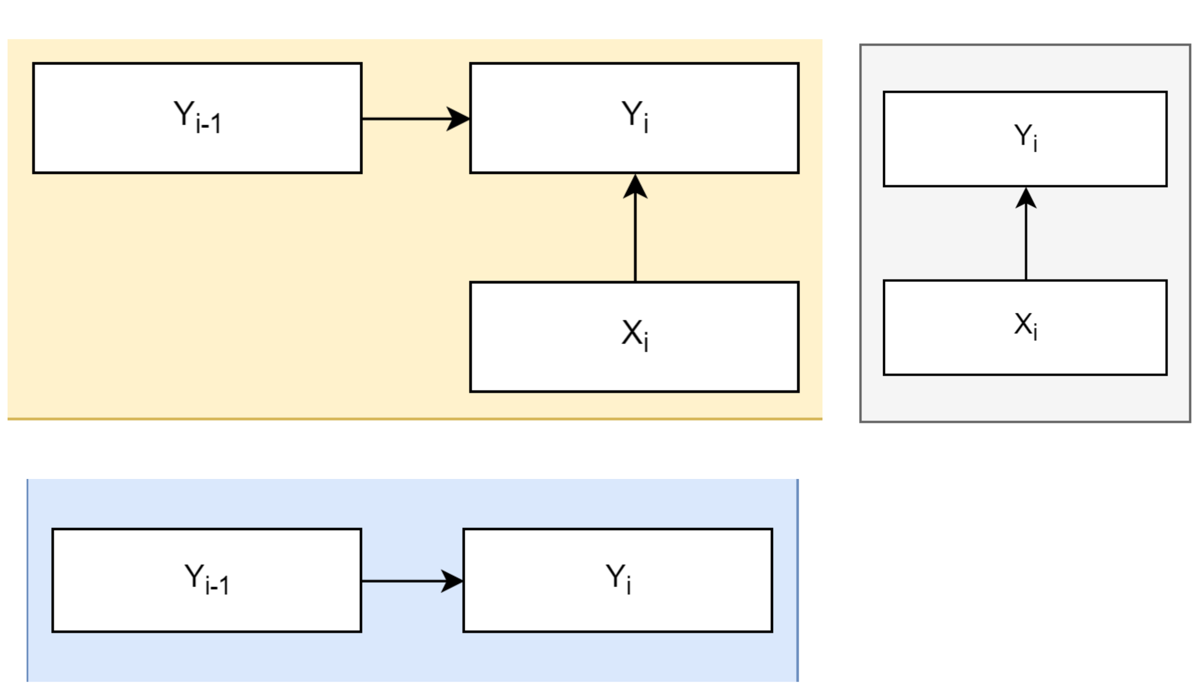
PREDICTED \ ACTUAL	Qty_s	Qty_e	Name	Unit	Comment
Qty_s	100.0%	0.0%	0.0%	0.0%	0.0%
Qty_e	0.0%	100.0%	0.0%	0.0%	0.0%
Name	0.0%	0.0%	92.7%	0.3%	0.6%
Unit	0.0%	0.0%	2.1%	89.3%	14.9%
Comment	0.0%	0.0%	5.2%	10.4%	84.5%

Given: [3, larg, onion, thinly, sliced]
Comment: [larg, thin, slice] Name: [onion]
QtyStart: 3.0 QtyEnd: 0.0
Unit: [nan]
HMM label: [qtys, comment, name, comment, comment]

Given: [1/4,cup, extra-virgin, oliv, oil]
Comment: [extra-virgin] Name: [oliv, oil]
QtyStart: 0.25 QtyEnd: 0.0
Unit: [cup]
HMM label: [qtys, unit, name, name, name]

STRUCTURED PERCEPTRON

We implement Structured Perceptron algorithm and compute accuracy. We do not remove stop words here and include them in our data. SP gives accuracy of **55.9%** in test data



Feature selection for HMM and SP

- HMM – Emission and Transition probability
- SP – Features include emission, transition and transition for observed emission

PREDICTED \ ACTUAL	Qty_s	Qty_e	Name	Unit	Comment
Qty_s	100.0%	0.0%	0.0%	0.0%	0.0%
Qty_e	0.0%	100.0%	0.0%	0.0%	0.0%
Name	0.0%	0.0%	90.9%	0.5%	0.4%
Unit	0.0%	0.0%	0.5%	80.2%	6.8%
Comment	0.0%	0.0%	8.5%	19.3%	92.8%

TEST DATA WORD LEVEL ACCURACY

Given: [2, ounc, cabbag, cut, in, 1/2-inch, slice, option]
Name: [cabbage]
Comment: [cut, in, 1/2-inch, slice, option]
QtyStart: 2.0, QtyEnd: 0.0
Unit: [ounce]
SP Label: [qtys, unit, name, comment, comment, comment comment, comment]

Given: [2, whole, dry, red, chili, like, thai, cayenne, or, arbol]
Comment: [whole, like, thai, cayenn, or, arbol]
Name: [dry, red, chili]
QtyStart: 2.0, QtyEnd: 0.0
Unit: [nan]
SP label: [qtys, comment, comment, comment, name, comment, comment, name, comment, comment]

CONCLUSION & FUTURE-WORK

We see from the weights that common transitions are scored more and selected more often. Certain confusables and noise in data-set can be removed to improve accuracy