Kansas State University Undergraduate Research Conference

Spring 2019

Identifying Sentences with Recipe information with Natural Language Processing

Erick Saenz-Gardea

Recommended Citation

Saenz-Gardea, Erick (2019). "Identifying Sentences with Recipe information with Natural Language Processing," *Kansas State University Undergraduate Research Conference*. https://newprairiepress.org/ksuugradresearch/2019/posters/47

Identifying Sentences with Recipe information with Natural Language Processing

Erick Saenz-Gardea and Dr. William Hsu

Department of Computer Science and College of Engineering, Kansas State University

1. BACKGROUND

Natural language processing (NLP) is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language. NLP software is used to translate text, identify spam in your email inbox, and identify relevant information online when using a search engine.

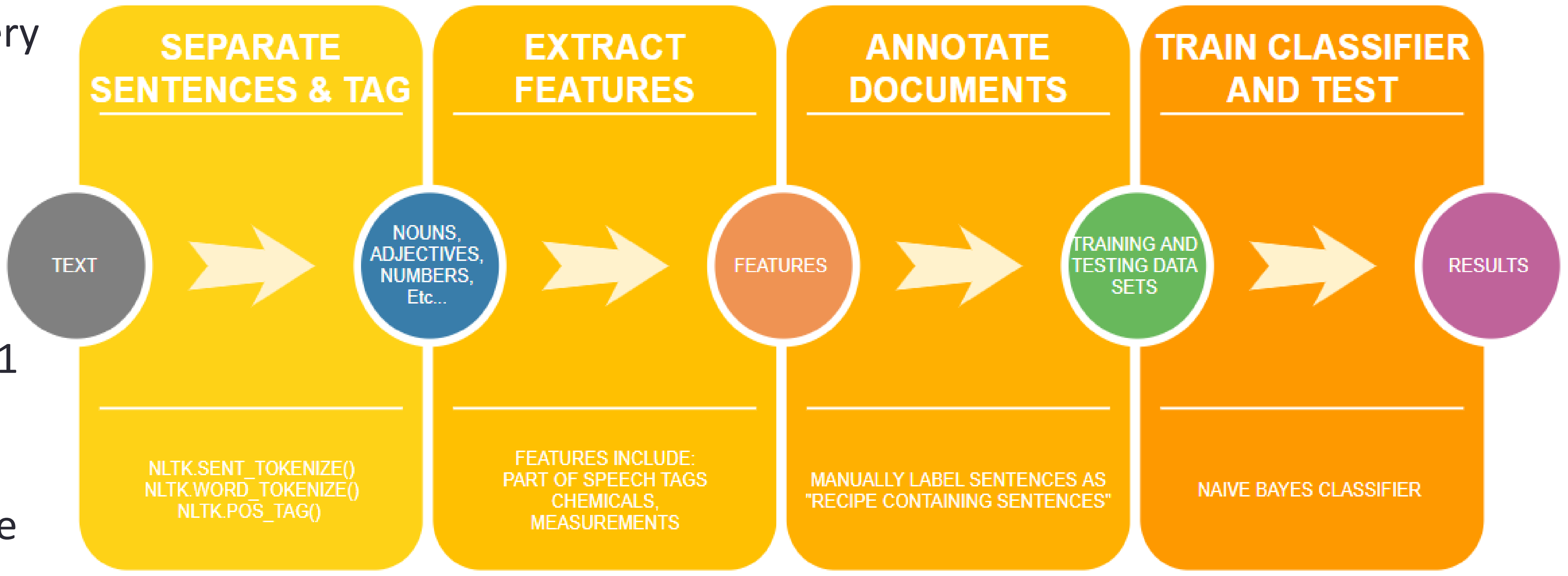
For a frame of reference, when reading text from another language, we may recognize letters, numbers, or punctuation, yet the underlying meaning is a complete mystery to us. With the help of NLP, computers can answer questions much more accurately, summarize entire documents, and quantify the connotation of an article.

The goal of this project is to determine whether a sentence contains a recipe step. We identify four features from the text: It's semantics, digits, chemicals, and wikification possibilities (in-text entities that are found on Wikipedia pages). These features are then used to train the Naïve Bayes Classifier. Afterwards the classifier was tested on documents within the same field.

2. METHODS

- Break the text into sentences and then extract every word in the sentence.
- Tag the words with its Part of Speech identifier
- Begin extracting features:
  1. Chemical names and formulas via PubChemPy API
  2. Digits (in terms of measurements such as 1 ml, 1 mg, and 1 kg)
  3. Wikification possibilities
- Manually Annotate the text and identify the recipe contain sentences for training/testing purposes.
- Train the Naïve Bayes Classifier.
- Test the Naïve Bayes Classifier.

- Create additional algorithms to create Quadgrams.
- Identify recipes via positioning.



SEPARATE SENTENCES & TAG

EXTRACT FEATURES

ANNOTATE DOCUMENTS

TRAIN CLASSIFIER AND TEST

TEXT

NOUNS, ADJECTIVES, NUMBERS, Etc...

FEATURES

TRAINING AND TESTING DATA SETS

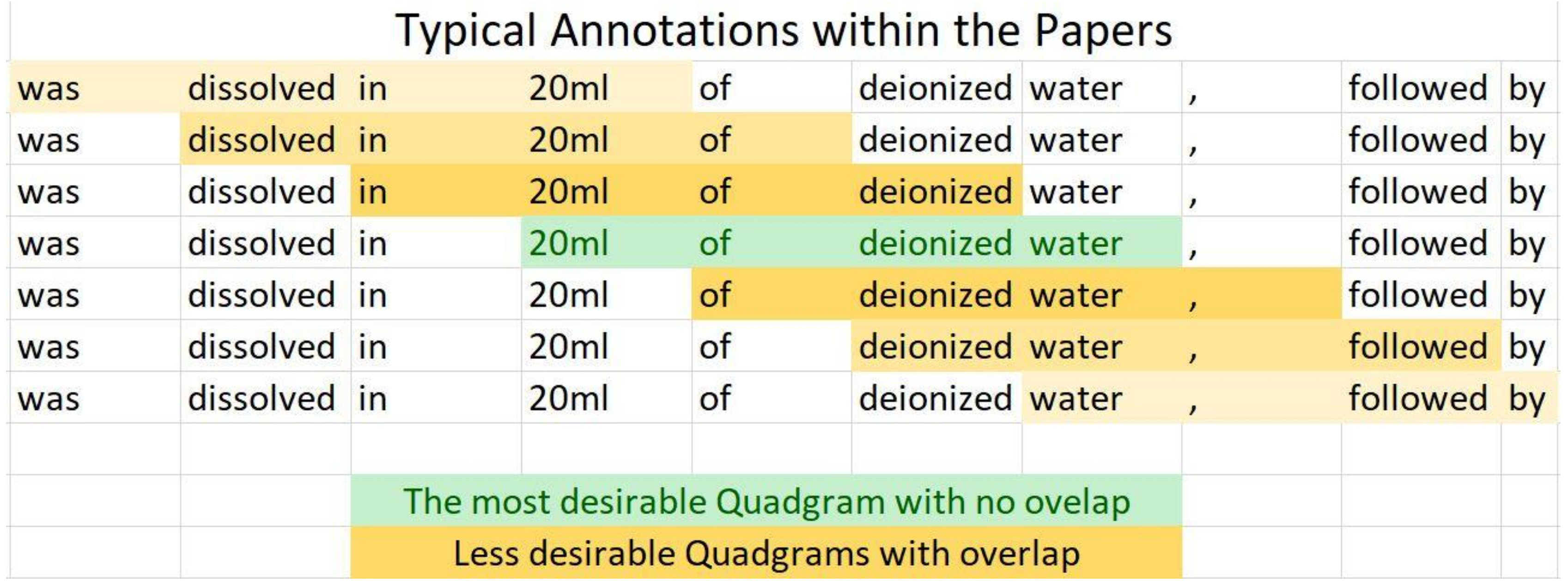
RESULTS

NLTK.SENT_TOKENIZE()
NLTK.WORD_TOKENIZE()
NLTK.POS_TAG()

FEATURES INCLUDE:
PART OF SPEECH TAGS
CHEMICALS,
MEASUREMENTS

MANUALLY LABEL SENTENCES AS "RECIPE CONTAINING SENTENCES"

NAIVE BAYES CLASSIFIER



Typical Annotations within the Papers

was	dissolved	in	20ml	of	deionized	water	,	followed	by
was	dissolved	in	20ml	of	deionized	water	,	followed	by
was	dissolved	in	20ml	of	deionized	water	,	followed	by
was	dissolved	in	20ml	of	deionized	water	,	followed	by
was	dissolved	in	20ml	of	deionized	water	,	followed	by
was	dissolved	in	20ml	of	deionized	water	,	followed	by
was	dissolved	in	20ml	of	deionized	water	,	followed	by
		The most desirable Quadgram with no ovelap							
		Less desirable Quadgrams with overlap							

3. RESULTS

Naïve Bayes Classifier Results

	Precision	Recall	F1-Score	Support
Positive	0.75	0.11	0.19	27
Negative	0.53	0.96	0.68	28
Average	0.64	0.55	0.44	55

- Positives are examples of sentences that contain recipe information
- Negatives are examples of sentences that do not contain recipe information
- The Support Column informs us of how many sentences were used for this project

Identifying Ingredients Results

	Precision	Recall	Total
Positives	1	0.64286	14

- All ingredients were correctly identified
- There exists redundant repetition in the output that was not accounted for.
- F1-Score is 0.7826

4. SUMMARY/INTERPRETATION

- Detecting positives worked 75% of the time.
- Collectively there are too many false positives being detected, henceforth the low recall rate of 11%
- There is no correlation between wikification possibilities and the recipe containing sentences.
- 75% of the sentences have wikification possibilities
- The average wikification possibilities are 3 per sentence
- Our first attempts at identifying ingredient only were fairly successful due to the high F1-Score

5. FUTURE WORK

For the future of this project we will look closer for the recipe components of the manufactured materials. This way, we may continue improving the previously developed algorithms to identify digits, chemicals, and semantics of the text. We will also use deep learning for labeling the role of an ingredient for recipe assembly. We should also continue to annotate more documents to see if the accuracy of this work improves. Typically, we see other projects of this nature use thousands of examples to train classifiers.

Alongside this we may also exclude the wikification feature due to it's limited utility in identifying the recipe steps. We may also expand the measurement category to include Moles, molarity, and acidity (PH).

6. ACKNOWLEDGMENTS

The Knowledge in Databases and Discovery (KDD) lab
led by Dr. Hsu
The Developing Scholars Program
The Ronald E. McNair Postbaccalaureate Achievement
Program