Web Scraping-Natural Language Processing (NLP) for Disease Outbreak Detection and Information Extraction

Yijun Wei, NISS-NASS
Luca Sartore, NISS-NASS
Nell Sedransk, NISS-NASS
Gavin Corral, USDA-NASS
Emilola Abayomi, USDA-NASS

Disclaimer

The findings and conclusions in this presentation are those of the author and should not be construed to represent any official USDA or U.S. Government determination or policy.

Outline

1. Background
2. Purpose
3. Approach:
    - Stage 1: Disease outbreak detection
    - Stage 2: Web scraping-Natural Language Processing (NLP) approach
        - ❑ Named Entity Recognition (NER) and Information Extraction (IE)
        - ❑ Stage 2.1: Hybrid approach
            - ➢ Definition
            - ➢ Case study
        - ❑ Stage 2.2 to improve Stage 2.1: Pre-training of Deep Bidirectional Transformers for Language Understanding (BERT)
            - ➢ Definition
            - ➢ Case study on NER
4. Conclusion
5. Potential for information

Outline

1. Background
2. Purpose
3. Approach:
    - Stage 1: Disease outbreak detection
    - Stage 2: Web scraping-Natural Language Processing (NLP) approach
        - ❑ Named Entity Recognition (NER) and Information Extraction (IE)
        - ❑ Stage 2.1: Hybrid approach
            - ➢ Definition
            - ➢ Case study
        - ❑ Stage 2.2 to improve Stage 2.1: Pre-training of Deep Bidirectional Transformers for Language Understanding (BERT)
            - ➢ Definition
            - ➢ Experiment on NER
4. Conclusion
5. Potential for the information

Background

1. NASS annually publishes estimates of US Hogs – Quarterly Hog Inventory National Numbers using statistical models
   - Models only capture equilibrium picture but cannot detect unusual pattern such as disease outbreak
2. The impact of disease can be substantial, but initial event can be local and/or small
3. Detection of the early stages is challenging
4. Current diagnostics lag the occurrence of the outbreak by a quarter
5. A preliminary study depended on the hybrid NLP approach with limited training data

Goals for two stages Web scraping-NLP approach

Stage 1 - Web scraping

- Goal: Hog disease outbreak rapid detection
- Application
  - ❑ Timely report for the disease outbreak

Stage 2 - Web scraping - NLP

- Goal: Information summary from scraped news
- Application of information summary
  - ❑ Geo-locating the outbreak
  - ❑ Better precision in predicting pattern and rate of spread of the disease
  - ❑ Documentation of the disease on each scale national-state-local with limited training data
  - ❑ Input to spatial-epidemic model
  - ❑ Data prepared for experts

Stage 1 - Web scraping
Hog Disease Outbreak Detection

Goal

- Detect whether the disease outbreak happens using web scraping

Tools

1. Web scraping in Swine Disease Global Surveillance Project (SDGSP)
   - ❑University of Minnesota Swine Center
   - ❑Monitors hog disease outbreaks on international scale
   - ❑Publishes reports every two weeks

African Swine Fever Outbreak: China

MONDAY, JANUARY 16, 2018 | BI-MONTHLY UPDATE

Download the report

Report highlights:

- China reports the first ASF outbreaks in Gansu and Ningxia provinces
- Multiple detections of ASF in products smuggled into Australia
- FMD outbreak in South Africa's FMD free zone

More Web scraping Sources

2. Disease Report Repositories
   - APHIS (USDA)

3. Media sources
   - News feeds – national, state, local
   - Extension service websites
   - Producers' organizations websites
   - Blogs

Outline

1. Background
2. Purpose
3. Approach:
   - Stage 1: Disease outbreak detection
   - Stage 2: Web scraping-Natural Language Processing (NLP) approach
     - ❑ Named Entity Recognition (NER) and Information Extraction (IE)
     - ❑ Stage 2.1: Hybrid approach
       - ➢ Definition
       - ➢ Case study
     - ❑ Stage 2.2 to improve Stage 2.1: Pre-training of Deep Bidirectional Transformers for Language Understanding (BERT)
       - ➢ Definition
       - ➢ Experiment on NER
4. Conclusion
5. Potential for the information

Stage 2: Web scraping-Natural Language Processing (NLP)

Step 1. Web scraping for news resources

- Multiple hog domain websites:
  - ❑ The Pig Site (https://thepigsite.com/)
  - ❑ National Hog Farms (https://www.nationalhogfarmer.com/)

Step 2. NLP for information summary

- Summarize information from the web-scraped news with the help from two NLP tasks:
  - ❑ Named Entity Recognition (NER)
  - ❑ Information Extraction (IE)

Named Entity Recognition (NER) and Information Extraction (IE)

1. Named entities (NE): definite noun phrases that refer to specific types of nouns
2. NER: identify all mention of NEs
3. IE: Identify structured relationships between NEs

Example:

The Ministry of Agriculture and Rural Affairs said the first outbreak is on a farm in the Xushui district of Baoding city

NER: Ministry of Agriculture and Rural Affairs, outbreak, Xushui district of Baoding city

IE: **Ministry of Agriculture and Rural Affairs said the first outbreak**

outbreak is on a farm in the Xushui district of Baoding city

Natural Language Processing (NLP) approaches overview

1. Rule-based
   - A hand-crafted system of rules based on linguistic structures that imitates the human
     - ❑ Eg: In NER, word shape feature, a word initialed with capitalized character, such as China, United States, etc … or a word with the format X.X.X, such as I.M.F
2. "Traditional" machine learning
   - Based on algorithms that learn to "understand" language without being explicitly programmed:
     - ❑ probabilistic modeling, likelihood maximization, and linear classifiers
       - ➢ Engineer features
       - ➢ Split the dataset into training data, and test data
       - ➢ Training a model on parameters, followed by testing on test data
       - ➢ Inference
3. Hybrid approach = 1+2

Stage 2.1: Hybrid approach

As soon as a hog disease outbreak is detected:

Step 1: Related news will be scraped from the website

Step 2: Information will be extracted from related news, using a hybrid approach

1. Normalize time
   - ❑ Different temporal formats transformed to a single form
2. Normalize word
   - ❑ Different word formats converted to a singular form
3. Keywords defining
   - ❑ Keywords are defined, eg: “outbreak”, “African Swine Fever”
4. Named Entity Recognition
   - ❑ Recognize the Named Entities
5. Information extraction
   - ❑ Extract pertinent information between Named entities

Application of the hybrid approach – Case study

Input, News item web-scraped from **The Pig Site**

'The Ministry of Agriculture and Rural Affairs said the first outbreak is on a farm in the Xushui district of Baoding city which has 5,600 hogs, some of which died because of the swine fever, though it did not provide a death toll.

The farm has been quarantined and the herd slaughtered, it added.

Application of the hybrid approach – Case study

Reuters reports that the second outbreak is in the remote Greater Khingan Mountains in Inner Mongolia, where 210 of the 222 wild boar raised on the farm died, the ministry said in a separate statement. The rest have been slaughtered, it said.

China has reported more than 100 cases of African swine fever in 27 provinces and regions since last August. The disease is deadly for pigs but does not harm humans.'

Case study specific steps

1. Normalize word:
   - Raised to raise
   - Slaughtered to slaughter
   - Quarantined to quarantine
2. Keyword define:
   - “outbreak”
3. Named Entity Recognition:
   - Ministry of Agriculture and Rural Affairs
   - Xushui district of Baoding city
   - Swine fever

Case study result

4. Information extraction, and result:

1. 'outbreak', Source: Ministry of Agriculture and Rural Affairs', Location: Xushui district of Baoding city', Stats: 5,600'

2. 'outbreak', Source: 'Reuters', Location: 'remote Greater Khingan Mountains in Inner Mongolia', Stats: '210 of the 222 died'

Hybrid approach drawbacks

1. Task specific for each disease and country
   - “Reporting it wouldn't have made a difference, he said, standing outside his farm in xijiahe, a village in China's Shandong province.”
2. Time consuming for coding rules
3. Corpus is not comprehensive
4. News from only a few hog news websites are scraped

Outline

3. Approach:
   - Stage 1: Disease outbreak detection
   - Stage 2: Web scraping-Natural Language Processing (NLP) approach
     - ❑ Named Entity Recognition (NER) and Information Extraction (IE)
     - ❑ Stage 2.1: Hybrid approach
       - ➢ Definition
       - ➢ Case study
     - ❑ Stage 2.2 to improve Stage 2.1: Pre-training of Deep Bidirectional Transformers for Language Understanding (BERT)
       - ➢ Definition
       - ➢ Experiment on NER
4. Conclusion
5. Potential for the information

NLP approaches overview- Continue

3. Hybrid approach = Rule-based + “Traditional” machine learning
4. Deep learning
   - Feature engineering is skipped, as networks will "learn" important features
   - Streams of raw parameters (words) without engineered features, are fed into networks
   - Large training corpus (dataset)
5. Semi-supervised deep learning
   - Only small number of training instances are available
   - Pre-trained on billion records corpus
   - Fine-tuned to domain adaption

(Jurafsky and Martin, 2014; Goldberg, 2017)

Stage 2.2: Semi-supervised deep learning network Overview

Designed to improve Stage 2.1, hybrid approach

1. Increase the number of scraped news sites
   - Streaming news API
     - ❑ Collect related news from hundreds of thousand web sources
2. Improve the efficiency of the hybrid approach with limited training data
   - Semi-supervised deep learning network
     - ❑ Pre-training of Deep Bidirectional Transformers for Language Understanding (BERT)

Pre-training of Deep Bidirectional Transformers for Language Understanding (Bert) Overview

1. One of the breakthrough in NLP in 2018, and outperforms all other models on various tasks
2. Pre-trained by Google AI Language on a large amount of text data, 3.3B words for 40 epochs
3. Can be fine-tuned on small data NLP task such as NER, and IE
4. Transfer learning from Human (also in computer vision)
5. Words are embedded based on both themselves and their neighborhood words
   - Word embedding: a word represented by a vector
     - Eg: run -> (0.2, 0.1, …, 0.3)

Case study

1. Fine-tune BERT to NER:
   - The Bert is pre-trained first
   - Adding a classification layer to task 2
2. News scraped from streaming news API
3. Corpora: 20 training text news + CoNLL03
4. Trained 5 epochs, and tested on 20 text news
5. Finished within 0.5 hour on a GTX 1070 GPU

Case study - Training dataset

Corpora (training dataset):

- CoNLL 2003: Dataset that contains 1,393 English news articles with annotated entities LOC (location), ORG (organization), PER (person) and MISC (miscellaneous)
- Text news: News article scraped from streaming news API using keywords “African Swine Fever”, and hand tagged the same as CoNLL 2003
- Using I (inside) - O (outside) - B (Begin), Each word is tagged with one of special chunk tags

Case study

Input

The Ministry of Agriculture and Rural Affairs said the first outbreak is on a farm in the Xushui district of Baoding city which has 5,600 hogs, some of which died because of the swine fever, though it did not provide a death toll.

Output

The Ministry of Agriculture and Rural Affairs said the first
O B-LOC I-LOC I-LOC I-LOC I-LOC I-LOC O O O
outbreak is on a farm in the Xushui district of Baoding
B-MISC O O O B-LOC I-LOC I-LOC I-LOC I-LOC I-LOC I-LOC
city which has 5,600 hogs, some of which died because of the
I-LOC O O B-MISC O O O O O O O
swine fever, though it did not provide a death toll.
I-MISC I-MISC O O O O O O O O

Case study
Evaluation metrics, Result

Confusion Matrix

	True condition		
Predicted Condition	Total population	Condition Positive	Condition Negative
	Predicted Positive	True Positive (TP)	False Positive (FP)
	Predicted Negative	False Negative (FN)	True Negative (TN)

F1 score: Precision*Recall/(Precision+Recall)
where: Precision = TP/(TP+FN)
Recall = TP/(TP+FP)

BERT Reached test F1 score 0.81

Case study - an example - Continued

	Confidence	Tag
Reuters	99%	B-ORG
reports	99%	O
that	99%	O
the	99%	O
second	99%	O
outbreak	99%	O
is	99%	O
In	99%	O

	Confidence	Tag
in	99%	O
the	99%	O
remote	99%	B-LOC
Greater	99%	I-LOC
Khingan	99%	I-LOC
Mountains	99%	I-LOC
in	99%	O
Inner	99%	B-LOC
Mongolia	99%	I-LOC
,	99%	O

Case study - an example

	Confidence	Tag
where	99%	O
210	99%	B-MISC
of	99%	O
the	99%	O
222	99%	B-MISC
wild	99%	O
boar	99%	O
raised	99%	O
on	99%	O
the	99%	O

	Confidence	Tag
farm	99%	O
died	99%	O
,	99%	O
the	99%	O
ministry	99%	O
said	99%	O
in	99%	O
a	99%	O
separate	99%	O
statement	99%	O
.	99%	O

Conclusion

1. Using BERT accelerates the process
2. More Corpora and training instances are needed for BERT to achieve better F1 score
3. BERT should be used for Information Extraction (IE) to summarize information from scraped news
4. The entire process is still in a very early stage, and more work should be done
5. More sources should be used for hog disease detection

Potential for information

Time and location of disease references:

1. Fine scale (state, county) incidence allowing spatial disease modeling and mapping
2. Time course of spread
3. External documentation confirming disease and response to outbreak
4. Data for other experts
5. Information to incorporate into the model system

Reference

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.".

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Goldberg, Y. (2017). Neural network methods for natural language processing. *Synthesis Lectures on Human Language Technologies*, *10*(1), 1-309.

Jurafsky, D., & Martin, J. H. (2014). *Speech and language processing* (Vol. 3). London: Pearson.

Sarawagi, S. (2008). Information extraction. *Foundations and Trends® in Databases*, *1*(3), 261-377.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

Wei, Y., Sartore, L., Miller, D., Abernethy, J., Toppin, K., & Hyman, M. (2018). Deep learning for Data Imputation and Calibration Wighting. In JSM Proceedings, Statistical Computing Section. Alexandria, VA: American Statistical Association.

Questions?