УДК 004.89+004.912

A TAXONOMY CREATION FOR AGRICULTURE USING CLASSICAL MACHINE LEARNING ALGORITHMS

M.E. PAULAVETS
Master student of the BSU,
Data Scientist EPAM Systems

J. PORCIELLO
The Associate Director of Research Data Engagement and Training in International Programs, Cornell University

Y.I. KIRYLLAU
Senior Data Scientist,
Project Manager EPAM Systems

S. EINARSON
Director of IT for International Programs, CALS, Cornell University

EPAM Systems, Republic of Belarus
Cornell University, the United States of America
Belarusian State University, Republic of Belarus
E-mail: maryia.paulavets@gmail.com

M.E. Paulavets
Graduated from the Belarusian State University. The master student of BSU in the specialty "Algorithms and systems for processing large volumes of information." Works at EPAM Systems as a Data Scientist. She conducts research in the field of natural language processing.

J. Porciello
The Associate Director of Research Data Engagement and Training in International Programs, College of Agriculture and Life Sciences, Cornell University. Director of TEEAL, a world agricultural science information resource. Program manager for training services for Research4Life, a United Nations public-private partnership. Evidence and communication leader in the Ceres2030 program.

Y.I. Kiryllau
Graduated from the Belarusian State University. Works at EPAM Systems as a Senior Data Scientist and Project Manager. Leads natural language processing projects.

S. Einarson
The director of transnational learning and head of information technology in CALS International Programs. In Asia and Africa, he contributes to e-learning and mobile solutions to promote curriculum and solve real-world problems.

Abstract. The Ceres2030 is an evidence and cost modeling program to support donor-decision making on high-impact interventions needed to end hunger and transform the lives of the world's poorest farmers (Sustainable Development Goal 2). Policy and decision-makers are interested in finding useful techniques and approaches to address urgent problems. Our goal was to automate the process of finding interventions, a colloquially used term, in articles and to make it easier for researchers and non-researchers search for scientific achievements. We used machine learning semantic models to generate a taxonomy of agricultural interventions and outcomes relevant to policy-makers. The intervention classifier was built with the help of classical machine learning algorithms, and our first results show the possibility of making use of even small datasets for natural language processing tasks.

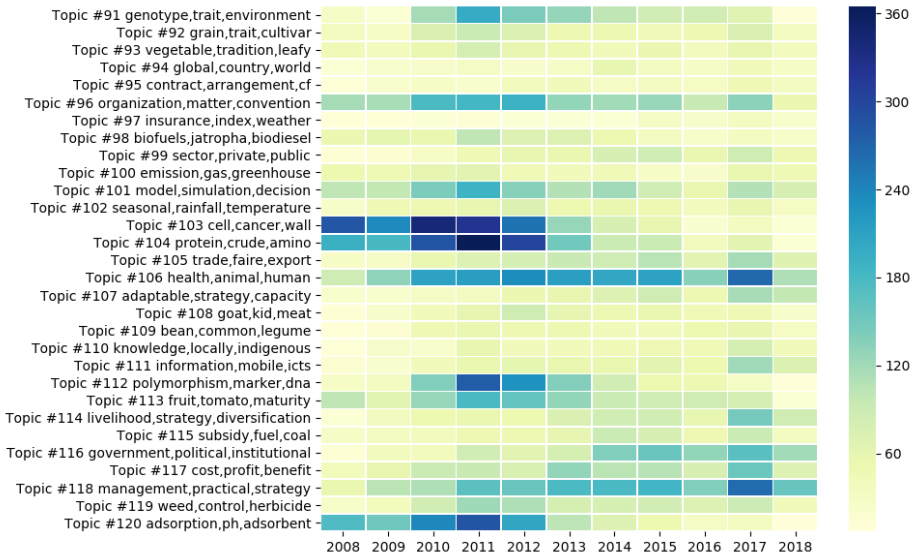
Keywords: machine learning, natural language processing, word embeddings, latent semantic analysis, taxonomy creation, agriculture policies, Word2vec, Hearst patterns

The CERES 2030 program took a mission to build consensus on the interventions needed to end hunger and transform the lives of the world's poorest farmers – while protecting the environment.

When researchers are looking for interventions and evidence of SDG indicators, they face challenges related to the volume of information and searching multiple databases. According to estimates from Science in 2013, a new scientific article is published every seven seconds. There are more than 30,000 journals publishing in the sciences, and for information related to SDGs, grey literature such as agency reports from the World Bank and FAO are also important resources. The process of articles investigation is quite time consuming and biased.

Our dataset consisted of 49910 articles' metadata from different scientific research sources. Metadata included a title, an abstract, author keywords, the year of publication, information about author and author affiliations. We didn't analyze the whole article, as it can be quite difficult to obtain access to the full publication, and we assumed that to a certain degree we can rely on an abstract and a title to derive useful information. The main theme of all these articles was agriculture and small-scale food producers – i.e., smallholder farmers. All the articles were published within the period of 2008-2019.

The first part of the articles' investigation was topic modeling. Most articles sources had only metadata analytics and relied on author keywords. But authors sometimes write too broad categories for keywords or don't specify at all; they also do not have a taxonomy to select from We decided to look at the topics generated by such machine learning algorithms as LDA and NMF [1]. Latent semantic analysis was performed over articles' titles and abstracts. Topics' number was chosen, so that we could see more specific topics. The bigger the number is, the narrower topics the model tends to have. For our purposes 150 topics were generated. We were particularly interested in topics changing in volume over time, as it can help researchers to see some information gaps and trends in research.



Picture 1. A research trend visualization of topics in agriculture.

One of the main problems of natural language processing is the possibility to express the same thing differently via rich language vocabulary. Researchers are constantly trying to keep up with a lot of synonyms to find relevant articles to find relevant articles using a 'keyword' search

method. - Machine learning does a good job in finding synonyms for the words using distributional methods. Word2Vec CBOW (continuous bag of words) model is trained to predict a word with the help of context words in the sentence [2]. The main advantage of this model is word embeddings derived by the model while training. The model tries to project the words onto smaller dimension space than original vocabulary's space. And due to reduced dimensionality, we can notice that similar by meaning words tend to be closer to each other. This approach is quite useful, but it has a disadvantage: the corpus for training should be big enough to get good results. The common practice to tackle such a problem is transfer learning. We used the Google News Word2Vec pre-trained model and trained our own model on top of it with the help of the Gensim library [3].

The Word2Vec word embeddings helped to identify synonyms for some necessary terms in agriculture. The results for word "intervention" can be seen on the table 1.

For researchers to discover - interventions in agriculture we first need a method to identify interventions outside of using keyword searching. We created a taxonomy of policy-relevant interventions via Hearst patterns extraction and solving classification problem.

Table 1

Table with synonyms for word "intervention"

Recommendation	Nutrition education
Action	Strategy
Project	Programme (all variations)
Development project	Measure
Targeting/Targeted	Effort
Entry point	Initiative
Assistance	Policy
Support	Outcome

The Hearst pattern-based approach is a standard way to approach unstructured and unrestricted volumes of texts that have no precoding associated with it to derive broad concepts (hypernyms) and narrow concepts (hyponyms) [4]. A hypernym is something more generic and broader whereas a hyponym is a specific instance of a hypernym. For example, an animal is hypernym, a cat is hyponym of an animal. Several examples of patterns: <hypernym> such as <hyponym> and/or/, <hyponym>, <hypernym> including <hyponym> and/or/, <hyponym> and many other patterns. Hearst patterns helped to discover words similar to intervention that researchers use in the literature. This increases the likelihood of finding relevant materials where researchers describe a specific intervention even when that word is not used.

The process of Hearst pattern extraction is described below.

Raw text is normalized, noun expressions are united with adjectives, participles and so on. NN is a noun expression. Example of Hearst pattern, coded in terms of pos tagging:

NN such as (NN and|or|, NN)*

Anthropometric measures (NN) indicated that use of spate irrigation (NN) did not have significant nutritional effects (NN), suggesting the need for nutrition-sensitive interventions (NN), such as nutrition education (NN) and awareness (NN) and multisectoral collaboration (NN).

This example shows that we have nutrition-sensitive interventions as hypernym and 3 hyponyms (nutrition education, awareness, multisectoral collaboration). After extraction we get 3 pairs (nutrition-sensitive interventions, nutrition education), (nutrition-sensitive interventions, awareness) and (nutrition-sensitive interventions, multisectoral collaboration), i.e. a pair is constructed as (hypernym, hyponym). We are interested only in those pairs which hypernym has either "intervention" word or "technology" word or their synonyms [5].

After extraction we performed several cleaning techniques, as far as this approach can detect too common words as narrow concept, so we removed pairs where hyponyms are too common

words, such as development, access and so on. Words for filtering are taken from 5000 common English words dictionary [6].

All these interventions can be split into 4 broad categories: technology, socioeconomic, ecosystem and miscellaneous interventions. We wanted to categorize each narrow concept from pairs to be one of the class described above. The dataset for training a classifier consisted of 1000 narrow concepts. While training we encountered the problem of unbalanced dataset, the class distribution is shown on the table 2. But the unbalanced was defeated to a certain extent due to model tuning.

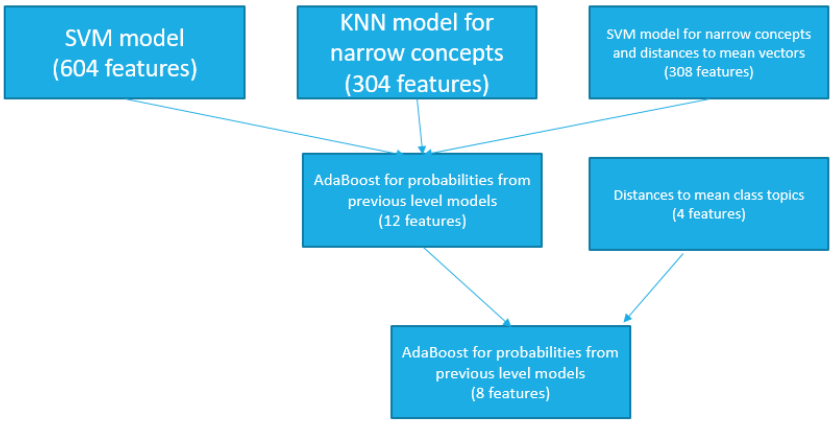
Table 2.

Dataset class distribution

Label	Number of narrow concepts
Technology intervention	550
Socioeconomic intervention	127
Ecosystem intervention	51
Miscellaneous interventions	272

The features for classification included a narrow concept word embedding (300-dimensional vector), all broad concepts, associated with a narrow concept – an average word embedding (300-dimensional vector), several numerical features, such as:

- Frequency of a narrow concept in articles
- Frequency of cooccurrence of a narrow concept together with word intervention and its synonyms in articles
- Weight of a narrow concept, evaluated on the top-5000 common English words dataset
- Weight of a narrow concept, evaluated on our full corpus (49910 articles)
- Distance of a narrow word embedding to the mean narrow concept word embedding of a particular class
- Distance of a broad word embedding to the mean broad concept word embedding of a particular class



Picture 2. The intervention classifier model

Our dataset is appropriately one thousand, so here we can't use neural networks, as far as neural networks are prone to overfitting on such small datasets. But classical machine learning algorithms can do quite a good job.

Support vector machines model works well with high dimensional data, as far as it also projects the data even onto higher dimension space itself. We tried to separate 4 classes by hyperplanes in the space and checked several models, but SVM with the radially based function gave better results (F1-measure - about 70%).

One of the well-known approaches to achieve higher accuracy in machine learning is to use model stacking and boosting. Boosting helps to make weak learners be strong learners together, but for better results models should be radically different, as far as they should investigate data from different sides and have uncorrelated errors.

K-nearest neighbors algorithm is the algorithm, which predicts a class for new points by majority of k neighbors votes. In our case, 5 neighbors gave sufficient results (F1 measure - about 65%).

It was a good idea to build SVM not only with word embeddings, but also distances to central word embedding of the particular class (F1 measure – about 67%).

All three models are joint by the model AdaBoost (boosting trees algorithms). AdaBoost's input is a 12d vector which is built from three 4d probability vectors. Each model gives a 4d vector where one vector component shows the probability of input data to belong to a particular class. For example, marker selected selection by first model is predicted as (0.63, 0.13, 0.09, 0.15), so the probability of class 1 is maximum and equals to 0.63, so it can be labeled as class 1– technology intervention. A 4d probability vector is produced by each of three models. And then AdaBoost builds decision trees of depth = 1 and tries to identify a label by all the probabilities the models gave it.

We added one more level of hierarchy in our model and the output of first AdaBoost model is concatenated with 4d vector of distances to mean topic embeddings derived by NMF for a particular class. And the last model which labels the narrow concept is AdaBoost with a 12d vector input. The whole model is shown on the picture 2.

Our final model has cross validation F1-measure up to 88%, and test F1-measure up to 84%. Some examples of the derived taxonomy:

- technology intervention: marker assisted selection
- socioeconomic intervention: agricultural extension service
- ecosystem intervention: agroforestry system
- miscellaneous intervention: association mapping

This taxonomy was used to label articles. Now researchers from the Ceres2030 program can find relevant articles with certain interventions easier and faster, not just looking through the whole dataset. Further research will be conducted to improve accuracy and investigate new possibilities to solve such kind of problems.

References

[1]. Hofmann T. Probabilistic latent semantic analysis //Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. – Morgan Kaufmann Publishers Inc., 1999. – P. 289-296.

[2]. Mikolov T. et al. Distributed representations of words and phrases and their compositionality //Advances in neural information processing systems. – 2013. – P. 3111-3119.

[3]. Gabrilovich E., Markovitch S. Wikipedia-based semantic interpretation for natural language processing //Journal of Artificial Intelligence Research. – 2009. – T. 34. – P. 443-498.

[4]. Hearst M. A. Automatic acquisition of hyponyms from large text corpora //Proceedings of the 14th conference on Computational linguistics-Volume 2. – Association for Computational Linguistics, 1992. – P. 539-545.

[5]. Word frequency data [Electronic resource]: URL: - https://www.wordfrequency.info/intro.asp (access date: 25.01.2019).

[6]. Gupta A. et al. Revisiting taxonomy induction over wikipedia //Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, December 11-17 2016. – 2016. – №. EPFL-CONF-227401. – P. 2300–2309.

[7]. Yang H., Callan J. A metric-based framework for automatic taxonomy induction //Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1. – Association for Computational Linguistics, 2009. – P. 271-279.

ИСПОЛЬЗОВАНИЕ КЛАССИЧЕСКИХ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПОСТРОЕНИЯ ТАКСОНОМИИ ДЛЯ СЕЛЛЬСКОХОЗЯЙСТВЕННОЙ ТЕМАТИКИ

М.Е. Павловец
Магистрант БГУ
Data Scientist
EPAM Systems

Дж. Порсиелло
Заместитель директора отдела исследования и подготовки данных по международным программам, Корнельский университет

Ю.И. Кириллов
Старший научный сотрудник, Руководитель проектов EPAM Systems

С. Эйнарсон
Директор по информационным технологиям для международных программ, CALS, Корнельский университет

EPAM Systems, Республика Беларусь
Корнельский университет, США
Белорусский государственный университет, Республика Беларусь
E-mail: maryia.paulavets@gmail.com

Аннотация. Ceres2030 - это программа моделирования фактических данных и затрат для поддержки принятия решений донорами относительно высокоэффективных улучшений, необходимых для прекращения голода и преобразования жизни самых бедных фермеров в мире (Цель 2 в области устойчивого развития). Лица, принимающие решения, заинтересованы в поиске полезных методов и подходов для решения насущных проблем. Наша цель состояла в том, чтобы автоматизировать процесс поиска улучшений, употребляемых в разговорной речи в статьях, и облегчить поиск научных достижений исследователями и не исследователями. Мы использовали семантические модели машинного обучения для создания таксономии сельскохозяйственных методов улучшения и результатов, имеющих отношение к ним. Классификатор методов и инноваций был построен с помощью классических алгоритмов машинного обучения, и наши начальные результаты показывают возможность использования даже небольших наборов данных для задач обработки естественного языка.

Ключевые слова: машинное обучение, обработка естественного языка, представления слов, латентный семантический анализ, создание таксономии, сельскохозяйственная политика, Word2vec, шаблоны Хёрст.