Usage Information for Real-Life Products

Bachelor project for 2022/2023
Main Advisor: Ralf Herbrich (Dept. of Artificial Intelligence and Sustainability)
Co-Advisor: Gerard de Melo (Dept. of Artificial Intelligence and Intelligent Systems)

Motivation. Today, products in (online) retail are predominantly structured via their category, price and brand. However, from a customer perspective, rather than knowing the appropriate category, they rather have a real-world task they would like to fulfill (e.g., travel) and it would be more desirable to search for the product using its intended usage. For example, a laptop may have uses that range from gaming to office work to school. Similarly, a pair of sports shoes may have different uses such as running, tennis, basketball, or soccer. Certain dresses could be used for formal occasions such as weddings, while others are more suitable for a casual lunch outing.

Usage information can be extremely useful for (1) helping customers in making purchase decisions, (2) improving relevance of search results, and (3) surfacing relevant search refinements so customers find the products they are looking for with a few clicks.

Solution Approach. In the proposed Bachelor projects, you must solve two related problems: First, use relation extraction techniques to identify candidate usage options for each product type/product. In a subsequent problem, you need to filter out the noisy candidate usage options and learn the mapping of products to usage options. In the first problem, we optimize for recall to ensure that we cover as many usage options as possible; while in the second phase, we focus on increasing the precision of usage options for all the products (individually).

Problem 1: Usage information for products appears in product titles and customer reviews. For example, today we find sentences in reviews such as “This is great for cutting hair”. We need to first generate a ground truth dataset of product review text and tagged usage options to use machine-learning methods for learning extraction techniques of usage options for products. Our focus will be on using crowdsourcing techniques and targeted crawls of the product catalog to generate a so-called “golden data set” of hand-labelled customer reviews. Like classical entity-extraction algorithms, in each sentence of a product review, the labelers need to assign each word to one of the four classes “START OF USAGE OPTION”, “MIDDLE OF USAGE OPTION”, “END OF USAGE OPTION” or “OTHER”. Then, a range of machine learning algorithms need to be learned and tuned and evaluated to learn to accurately predict one of these four classes (though modelling the sequence of START, MIDDLE and END appropriately may turn out advantageous). Once we have learned such a classifier, we will use it to predict labels of words in customer reviews. We will tune these classifiers to fine-tune the classifiers to give a large recall of all possible usage options across a range of customer reviews.

Problem 2: Once we have extracted a comprehensive candidate set of usage options, we will reuse the same “golden data set” to learn a function that maps a given review for a product to all the possible usage options. However, in contrast to Problem 1, we focus here on co-agreement between usage options across customer reviews and accuracy of the prediction (rather than recall). We may have to use a clustering algorithm such as k-means or spectral clustering to cluster the various usage options to optimize precision.

Data & Technologies. In this project, we will use the Amazon Customer Reviews dataset (publicly released at https://s3.amazonaws.com/amazon-reviews-pds/readme.html). It consists of over 130+ million customer reviews collected in a period of over two decades since the first review in 1995. The dataset is available in English, German, French and Japanese. For analysis, we will make use of Python and PyTorch. One starting point of algorithms will be Transformer models. For the crowdsourcing, we will consider Amazon Mechanical Turk and Amazon SageMaker GroundTruth.

Project Partner. Amazon Research in Berlin develops novel methods in machine learning (ML) and data science. The partner team, International Retail Machine Learning, is focusing specifically on applications of ML to problems in online retail to improve accuracy of online search and recommendation. All methods developed by the team will be directly applicable to production systems serving millions of customers daily. More information about the partner team can be found at https://www.amazon.science/ or email ralf.herbrich@hpi.de. We will have bi-weekly meeting with the project partner to both receive feedback and present project progress.