氏　　名	PAWEL CEZARY LEMPA
授与学位	博士（工学）
学位記番号	博甲第170号
学位授与年月日	平成30年9月10日
学位授与の要件	学位規則第4条第1項
学位論文題目	Development of Optimization Method with the Use of Genetic Algorithms for Natural Language and Related Models （遺伝的アルゴリズムを用いた自然言語とその関連モデルの最適化手法の開発）
論文審査委員	主査　准教授　桝井　文人 准教授　ウラ　シャリフ 准教授　プタシンスキ　ミハウ　エドムンド 教授　榮坂　俊雄 教授　平山　浩一

学位論文内容の要旨

Language models are an indispensable element of Natural Language Processing (NLP) research. They are used in machine translation, speech recognition, part-of-speech tagging, handwriting recognition, syntactic parsing, information retrieval and others. In short, language models are probability distributions over sequences of words. There are countless numbers of NLP solutions, algorithms and programs applying language models in specific tasks. Unfortunately, often these are not optimized, but rely on default, most commonly used sets of parameters. For example, many of them use numerous objective functions with different variables but without proper weights applied to them. Users usually set these variables themselves, which causes the results not to exceed a certain mediocre level. In case of small number of variables, users can adjust them manually, but optimization of objective functions with massive number of variables, especially multi-objective functions is difficult and time consuming. This was the motivation to propose an application of a Genetic Algorithms (GAs) to optimize the weighting process. GAs are subset of Evolutionary Algorithms (EAs), inspired by the process of natural selection known from nature. They use bio-inspired operators such as selection, crossover and mutation to generate solutions for optimization and search problems. This way GAs represent randomized heuristic search strategies simulating natural selection process, where the population is composed of candidate solutions. They are focused on evolving a population from which strong and diverse candidates can emerge via mutation and crossover (mating). There exist different types of GAs, moreover the same type of GA can bring different quality of solutions, depending on multiple variables, which include starting population, number of generations or fitness function. Finding the best starting parameters and type of GA the most appropriate for a given optimization problem is a next challenge. For that reason, I created a library that automatically applies multiple types of GAs in optimization purposes.

The library was created in C++ language, with the use of .NET environment. Its main goal is to be used with different secondary programs and applications, without significant interfering in the original structure of the solution. Basic function of library allows the use of several different kinds of GAs like: Simple GA, Uniform Crossover GA, n-point Crossover GA, GA with sexual selection, GA with chromosome aging and so forth. User can freely define starting parameters

for GA including: population size, starting population, number of generations, type of mutation and crossover. Advanced functions of the library allow the use of multithreaded processing for running several GAs in the same time. Basic option of multithreading runs the same type of GA with different starting parameters, advanced version allows to exchange information between different threads every set number of generations. In case of large number of variables to compute, it is also possible to separate a mutation and crossover for several threads running at the same time.

The most important functionality of the library is its easy adjustability in optimization of different kinds of applications. The library is used to run the original program in every generation of GA with new weights for variables generated from natural selection. Time of program running is closely related with original program processing time. It depends on the type of original solution and the time of processing one generation is similar to one run of the optimized program.

During creating and testing the library, numerous experiments have been carried out. In preliminary experiments the library was used for optimization of construction of mechanical elements. Later the application was tested on natural language processing and related solutions. One part of the research was optimizing Quantitative Learner's Motivation Model. The goal of this experiment was to optimize the formula for prediction of learning motivation by means of different weights for three values: interest, usefulness in the future and satisfaction. For this optimization, an application in C# using GA library was created. Data sets for the experiments were acquired from questionnaires enquiring about the above three elements in actual university classes. The results of the experiment showed improvement in the estimation of student's learning motivation up to over 17 percentage points of F-score.

The final experiment aimed to optimize the implementation of Support Vector Machines (SVMs) for the problem of pattern recognition in natural language data. SVMs are a machine learning algorithm based on statistical learning theory. They are applied to large number of real-world applications, such as text categorization, hand-written character recognition, etc. Original program was created in C++. For this application numerous different types of GAs were tested with different number of generations, weight range and starting parameters. Optimization was successful, with different scale of improvement based on previously mentioned conditions, with the highest achieved improvement of over 6 percentage points of recall comparing to baseline and reaching 78%. All experiments data are included in this work.

論文審査結果の要旨

遺伝的アルゴリズム（GA）は、計算量や局所解が懸念されるテーマに対して効果が期待できるメタヒューリスティックアルゴリズムであり、解候補の表現を工夫することによって組み合わせ最適化問題やNP困難問題などにも適用可能な利点がある。しかしながら、自然言語処理に対するGA応用を報告した先行研究は極めて限定的であり、その有効性について深く議論されていない状況にある。

本論文では、統計的言語モデルの最適化プロセスにおいてGAを活用することを提案している。具体的には、有害表現検出を目的とした言語モデル構築におけるパラメータ組み合わせの最適化において、比較的大量のデータが存在しない場合であっても効率よく最適解を導出できることを実験的に明らかにした。従来、有害性を高める語やフレーズのトレンドは時間と共に変化するため常に十分な訓練データを確保することが困難であったが、本研究によってこれに起因する学習効率の問題を大きく軽減できる。

これを要するに、著者は、自然言語処理において高度かつ難解とされる言語モデル最適化について有効な計算処理機構に関する新知見を得たものであり、自然言語処理が処理対象とし得る文書の範囲拡大に対して貢献するところ大なるものがある。よって著者は、北見工業大学博士（工学）の学位を授与される資格があるものと認める。