Streszczenie

Niniejsza praca skupia się na praktycznych zastosowaniach neurokognitywnych inspiracji w dziedzinie przetwarzania języka naturalnego. Za przykład posłużyła kategoryzacja tekstu z rozchodzącą się po sieci semantycznej aktywacją. To przedsięwzięcie nie należało do łatwych — podobne prace powstały w przeszłości, jednak bez większych sukcesów. Po wielu latach badań, zorientowano się, że problem tkwi w sieciach semantycznych, które zawierają duże ilości sprzecznej oraz niezwiązanej z tematem wiedzy.

Zjawisko to znane jest jako problem wielokrotnego dziedziczenia i może być obecne nawet w małych sieciach semantycznych. Celem będzie ukazanie, że prawidłowe zastosowanie rozchodzenia się aktywacji w sieci semantycznej wymaga **nie-semantycznego** hamowania. Nauki neurokognitywne uczą, że pacjentów z zaburzeniem myśli typu rozkojarzeniowego charakteryzuje nieprawidłowo funkcjonujące hamowanie w pamięci semantycznej. Powyższe stwierdzenie posłuży za odpowiednią analogię do tego, co się dzieje podczas kategoryzacji tekstu przy użyciu niezmodyfikowanej sieci semantycznej.

Zaproponowane zostaną dwa heurystyczne rozwiązania problemu wielokrotnego dziedziczenia: rozwiązanie, które używa ośmiu **nie-semantycznych** typów przycinania sieci semantycznych oraz rozwiązanie, które używa dwunastu **nie-semantycznych** typów przycinania sieci semantycznych. Następnie porówna się te podejścia do rozwiązania, które stosuje **semantyczne** relacje do przycinania oraz do rozwiązania, które nie używa rozchodzenia się aktywacji w sieci semantycznej. Nadto, przedstawione będą wybrane metody ewaluacji każdego z rozwiązań poza standardową miarą $F1$.

Zaprezentowane w pracy wyniki automatycznej kategoryzacji na zbiorze "Heart Diseases" opracowane zostały przez grupę ze szkoły medycznej OHSUMED. Jako sieci semantycznej użyto Unified Medical Language System Metathesaurus przygotowanej przez Amerykański Instytut Zdrowia. Zastosowanie wizualizacji oraz miar rozkładu, posłuży do przeprowadzenia porównania wewnątrz pojedynczego rozwiązania problemu hamowania, porównania pomiędzy różnymi rozwiązaniami oraz dokonania porównania z pracami innych grup naukowych. Dotychczas nikomu nie udało się pokazać, że kategoryzacja tekstu z wnioskowaniem poprzez aktywację zewnętrznej wiedzy może działać tak dobrze, jak kategoryzacja używająca słów oraz selekcji i ważenia cech. Ponadto zaoferowano przypuszczenie, które pozwoli wyjaśnić zmiany w standardowej mierze $F1$ oraz zmniejszyć ciężar obliczeniowy związany z wyszukiwaniem najlepszej metody hamowania rozchodzącej się aktywacji.

Na zakończenie wyciągnięte będą wnioski na temat praktyczności powyższego podejścia do przetwarzania tekstów medycznych i przedstawione zostaną dwie propozycje, aby **niesemantyczne** przycinanie grafu nazwać **pragmatycznym** przycinaniem oraz semantyczne relacje były używane celem gromadzenia wiedzy, a nie wnioskowania. W zamian, **pragmatyczne** relacje powinny być używane do wnioskowania oraz hamowania zbędnej wiedzy.

Abstract

This work focuses on practical applications of neurocognitive inspirations in the area of natural language processing. We applied text categorization with spreading activation over a large semantic network, which was extremely challenging. As we describe, there have been many similar but unsuccessful attempts in the past. Over our years of research, we have discovered that the problem lies in the semantic networks that contain large quantities of contradictory or irrelevant knowledge.

This phenomenon is known as the multiple inheritance problem. It can be present even in small semantic networks. We demonstrate that successful application of text categorization with spreading activation requires **non-semantic** inhibition. Neurocognitive sciences teach us that patients with formal thought disorder cannot perform certain cognitive tasks because they lack inhibition in their semantic memory. We maintain that this is a useful analogy to text categorization with unmodified semantic networks.

Using this neurocognitive lesson, we propose two heuristic solutions to the multiple inheritance problem: the eight **non-semantic** edge types pruning and the twelve **non-semantic** edge types pruning. These two solutions are contrasted to an approach that uses spreading activation with **semantic** edge types pruning and to an approach that does not use spreading activation at all. Further, three additional evaluation techniques of the recommended solutions are suggested that go beyond the typical $F1$ performance measure.

The results of the text categorization experiments are presented on the "Heart Diseases" subset of the OHSUMED document collection using the Unified Medical Language System Metathesaurus as the large-scale semantic network. Various visualizations and summary statistics were used to compare the **intra-model**, the **inter-model**, and the **external model**. To our knowledge, this is the first large-scale system to be proposed that uses conceptual text representation with inference to yield results on par with contemporary systems that use simple n-gram text representation with feature weighting and selection. Furthermore, we provide a conjecture that gives reasons for $F1$ performance changes and that may decrease the computational burden of the complex task of reasoning with **non-semantic** inhibition.

We conclude that the **non-semantic** edge types pruning is, in fact, **pragmatic** edge type pruning. The practicality of **non-semantic** edge types comes from their ability to increase or decrease text categorization performance. This is something that the **semantic** relationships cannot accomplish because they are designed to accumulate knowledge, not to use it.