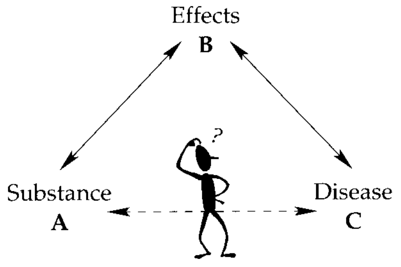
Literature-based discovery in biomedicine

Weeber, Marc

Summary

The research reported in this thesis is an example of biomedical research that does not take place in the traditional laboratory nor in a clinical setting. New knowledge does not only originate from test tubes (in vitro) or from patients (in vivo) but also from the interaction between the researcher and the computer (in silico). The premise of our research is that, nowadays, there is too much scientific knowledge to be captured and combined by a single human being. We postulate that the computer can assist the researcher in looking for knowledge outside his or her scientific domain. For instance, consider the knowledge on Raynaud's disease as available in 1985. Patients with this disease suffer from intermittent blood flow in the extremities such as fingers, toes, and ears. The blood viscosity in these patients was known to be elevated. Moreover, blood platelets of these patients were found to aggregate easily. In 1985, it was also known that the ingredients of fish oil, omega-3 fatty acids, induce a decrease of blood viscosity and platelet aggregation. However, this knowledge was not shared by the researchers involved in treating patients with Raynaud's disease.

In 1986, Don R. Swanson, a professor at the University of Chicago, formulated the hypothesis that patients with Raynaud's disease may benefit from fish oil, based on an extensive survey of the biomedical literature (Swanson, 1986). Clinical studies conducted after 1986 corroborated this hypothesis. Swanson has described this kind of scientific discoveries by a simple *ABC*-model:



There is a scientific discipline that has generated the knowledge that a substance *A* induces certain effects *B* in the body. Another discipline has the knowledge that manipulating effects *B* may help treating patients with disease *C*. However, nobody has made the direct connection between *A* and *C* yet.

Using this model of *connecting disconnected disciplines*, Swanson has made several discoveries in the biomedical domain since 1986. For this, he uses a computer analysis of MEDLINE titles. MEDLINE is the most comprehensive bibliographical biomedical database that incorporates information of over ten million scientific publications. Swanson's most famous discovery, a relationship between magnesium deficiency and migraine, is common knowledge at this moment. In 1988, Swanson found eleven implicit connections, or intermediate *B*-effects, between migraine and magnesium deficiency in the biomedical literature that no one had noticed before (Swanson, 1988).

The research by both Rein Vos (1991) and Floor Rikken (1998) uses a similar model. Instead of Swanson's interests in dietary factors, Vos and Rikken are interested in drugs (*A*) and their side-effects (*B*). Most drugs have a broad spectrum of effects in the human body, being favorable (effect) or not (side-effect). The (in)favorability of a certain effect is context-dependent (Rikken, 1998). For instance, increased hair growth by the antihypertensive drug minoxidil may be a side-effect, but for treating baldness it may be a favorable effect. **Chapter 6** provides more examples of side-effects that have become a favorable effect in different contexts.

The vast extent of the biomedical scientific knowledge is the cause that there are isolated, i.e. bibliographically not connected, disciplines. By connecting these disciplines, new scientific discoveries are possible. Ironically, the vastness of knowledge also provides the challenge how to find the proverbial needle in the hay stack. Following Swanson, we use MEDLINE as the written format of the biomedical knowledge domain. This means that the discoveries are hidden in the ten million titles and abstracts. In this thesis, we show that advanced computational linguistic analyses are able to reduce the staggering information space to a humanly manageable size. Besides simulating Swanson's discoveries, we have found new potentially therapeutic applications for the drug thalidomide.

The introductory **Chapter 1** shows that Vos' and Swanson's discovery models are identical. This implies that we may apply Swanson's analytical approach to Vos' and Rikken's research interests, drugs and adverse drug reactions or side-effects. In **Chapter 2** and **Chapter 3**, we therefore explore several techniques to identify side-effects in MEDLINE abstracts. For this, we compiled a collection, or corpus, of 539 abstracts that, among others, discuss side-effects of the angiotensin converting enzyme (ACE) inhibitors captopril and enalapril. We asked two pharmacologists to indicate which words are related to the side-effects of these two cardiovascular drugs. We used the resulting two word lists as a golden standard to compare our computational linguistic techniques with.

Chapter 2 describes our experiments with a standard computational linguistic technique of computing the association between words. We use two statistical tests to determine the significance of this relation: Fisher's exact test and the log-likelihood ratio. We observed a regularity that has not been known to date. This regularity is characterized by a saw-tooth-shaped pattern in a graph that plots the number of significant words. It turns out that the pattern originates from the fact that the lowest-frequency words (occurring only 1, 2, 3, or 4 times in the corpus) loose significance at predetermined moments. Since there are many of these lowest-frequency words, the effect is dramatic. Usually, these words are discarded from the analysis a priori, however, we show that the lowest-frequency words are as important as the higher ones. This holds true for both our pharmaceutical application and a totally different linguistic application. We therefore present an adaptation of the standard statistical analysis that adjusts for the phenomena observed.

Chapter 3 provides the results of a comparison of different techniques that identify side-effect-related words in the corpus. The rationale is that these words possess certain linguistic characteristics that are typical. There are many ways to define and compute characteristics, we explored three of them. First, we compute the association between the word "side-effect" and each other word in the corpus. The higher the association, the more likely it is that a word is related to side-effects of drugs. For the second technique, we compute seven information theoretic measures. Applying a classification tree algorithm, we try to find general characteristics or rules in these measures that are typical for side-effects. In the third technique, memory-based learning, we use contextual information. The assumption is that a word's context provides information on its relation with side-effects. We define context as two words to the right and two words to the left of a specific word.

Each technique results in a list of words that are related to side-effects. We compared these lists to the lists of the pharmacological experts. It turns out that the word association technique performs worst in terms of sensitivity and specificity. The classification tree analysis is able to extract robust classification rules, however, the memory-based learning algorithm performs best with an average sensitivity of 0.36 and an average specificity of 0.99.

One of the more interesting observations in **Chapter 3** is that representing pharmacological knowledge by single words is not very informative. We therefore use biomedical, often multi word, concepts in the remainder of the thesis. The concept *High blood pressure*, for instance, is more informative than the three single words *high, blood* and *pressure*. We use the concepts that are available in the Unified Medical Language System (UMLS) Metathesaurus, the largest collection of biomedical concepts with over 730,000 concepts in the 2000 version. The UMLS concepts have been categorized by so-called *semantic types*. Each concept has been assigned at least one out of the 144 possible types. The concept *High blood pressure*, for example, has been assigned the semantic type of **Sign or symptom** and the concept *Fish oils* has been assigned **Lipid**. Using the UMLS concepts as the units of analysis, we have developed, validated, and employed a literature-based discovery support system in **Chapter 4, 5** and **6**.

Chapter 4 describes the architecture of the *DAD*-system, an Internet-based computer system that provides the biomedical researcher with a tool to generate and test new hypotheses based on the scientific literature. The acronym *DAD* expands to *D*rug – *A*dverse drug reaction – *D*isease. The system uses MetaMap, a computer program that employs advanced *Natural Language Processing* techniques to identify UMLS concepts in MEDLINE titles and abstracts. We use the *semantic types* as a filter to only select the most interesting concepts from the texts. For instance, in our quest for new therapeutic applications for drugs, we are interested in concepts representing diseases. In this case, we only select the concepts that have been assigned the semantic type of **Disease or syndrome**. All other concepts are not interesting and are therefore discarded.

Both the literature on these kind of discovery systems and our own research show that filtering as described above is absolutely necessary to select the few potentially interesting concepts from many thousands of titles and abstracts. When evaluating the filtered concepts, the user is provided with the original context of the concepts, i.e., the sentences in which they occur. The user only has to study a mere hundred, very relevant, sentences. The human researcher's task is feasible now.

In **Chapter 5**, we simulate Swanson's two most well-known literature-based discoveries. First, we show the linking of Raynaud's disease and (the ingredients of) fish oil through different, indirect pathways (Swanson, 1986). Subsequently, we simulate the discovery of the relationship between magnesium deficiency and migraine (Swanson, 1988). We employ a two-step process. The first step, the greatest challenge, is to reach, from starting point *A* through intermediate step *B* the end point *C*. Beforehand, it is unknown where *C* is located in the information space of the ten million titles and abstracts. The *DAD*-system is able to distinguish the clearest *ABC*-pathways in this space. Once a connection between *A* and *C* has been established, the second step of elaborating, or testing, the link may be employed. Using this second step, we find the additional, less clear pathways between *A* and *C* as well. The fact that we use identical system parameter settings and semantic filters indicates the robustness of our approach.

In **Chapter 6**, we use the *DAD*-system for creating truly new scientific literature-based hypotheses. We use the drug thalidomide as a case study. In 1961, only two years after its introduction as a sedative, thalidomide was retreated from the market because of birth defects such as limb deformation in children from women using this drug. Because thalidomide has a broad spectrum of pharmacologic effects, it has continuously been investigated in medical experiments. In 1998, this led to the new registration for treating erythema nodosum leprosum, a skin disease manifest in leprosy patients.

The observation that a forty year old drug is used for a new application led us to a systematic analysis of the literature in order to find other new and unknown therapeutic applications. Using the *DAD*-system, we analyzed more that 60,000 titles and abstracts from MEDLINE in a limited amount of time. The powerful semantic filters reduce the huge amount of information to human manageable proportions. Based on these analyses, we conjecture that patients with myasthenia gravis, sialadenitis, acute pancreatitis, chronic hepatitis C or helicobacter pylori induced gastritis may benefit from thalidomide.

The indirect (*B*) connection between these diseases and thalidomide is an immunologic reaction: The balance in differentiation of the so-called immune T-helper cell. This cell can differentiate to T-helper 1 (Th1) or T-helper 2 (Th2) cells. The aforementioned diseases, and possibly also other autoimmune disorders, are characterized by an overdifferentiation towards Th1 cells. Thalidomide has the effect of inhibiting this process by differentiating towards Th2.

Chapter 6 shows that our computer-directed approach to discovery successfully generates new biomedical knowledge. We therefore explore the process of the act of making scientific literature-based discoveries in **Chapter 7**. It turns out to be a close collaboration between the computer, the domain expert, an immunologist in the thalidomide case, and the information specialist. Each of these players has to be an expert in its own field. The computer analyses are computationally intensive and storage demanding, the domain expert must be able to interpret the validity and value of the generated hypotheses, and the information specialist has to know the (technical) possibilities of information sources and retrieval systems. Interdisciplinearity is both the basis of the discovery model and the environment necessary for successful practical performance. We assume that computer-assisted literature-based scientific discovery systems as developed in this thesis may successfully be integrated into the practice of biomedical research.