FINAL REPORT

Measuring Semantic Relatedness using a Medical Taxonomy

by

Siddharth Patwardhan

August 2003

A report describing the research work
carried out at the Mayo Clinic in Rochester
as part of an internship in the summer of 2003

Department of Computer Science
University of Minnesota
Duluth, Minnesota 55812
U.S.A.

Acknowledgments

I would like to take this opportunity to thank a number of people, without whose support and encouragement this internship would not have been possible. Firstly, I would like to thank Dr. Serguei Pakhomov, who had the vision and was so supportive of the work. I would like to thank Ted Pedersen for his valuable advice throughout. I also would like to thank my colleagues at the Mayo Clinic – Bridget Thomson McInnes, Guergana Savova and Dara Becker – for their comments and suggestions. Finally, I would like to thank Mayo Clinic for giving me this wonderful opportunity.

Contents

1 The Company **2**

1.1 Mayo Clinic 2

1.2 Division of Medical Informatics Research 2

1.3 History of Medical Informatics Research at Mayo 3

2 The Project **4**

2.1 Previous Work 4

2.2 Medical Resources 5

2.3 Using a Medical Taxonomy to Measure Semantic Relatedness 5

3 Tools and the Environment **7**

3.1 Hardware and Software 7

3.2 Work Environment 7

4 Conclusions **8**

Abstract

Semantic Relatedness of words is a an intriguing problem in Natural Language Processing. Most humans agree on the relatedness of most word pairs. For example, if one said that “pencil” is more related to “paper” than it is to “boat”, there would be little opposition to this fact. Work by Budanitsky and Hirst (2001), Patwardhan and Pedersen (2003) attempt to compare a number of automatic measures of semantic relatedness. The experiments show that a number of the measures correspond closely to human perception of relatedness. The research also shows that these measures perform quite well in Natural Language Processing tasks.

The measures that were compared extensively use WordNet, a semantic network of real world concepts. A number of semantic networks and taxonomies, such as SNOMED CT and MeSH, exist in the Medical Informatics domain. Replacing WordNet with one of these would enable us to measure the semantic relatedness of medical concepts. Rich corpora of patient data is also available at large hospitals, which can be used to enhance the performance of these measures.

The research carried out at the Mayo Clinic, during this internship, attempts to implement various measures of relatedness using a semantic network of medical concepts called SNOMED CT. Initial experiments were conducted to test the performance of these measures in this specialized domain. The results showed some promise. Based on their performance, these measures may be applied to other tasks in the medical informatics domain – such as clustering of patient data, building new ontologies, etc.

1 The Company

1.1 Mayo Clinic

Mayo Clinic is located in Rochester, Minnesota, about an hours drive from the Twin Cities. It is one of the major employers (apart from IBM) in the small town. Mayo Clinic is ranked among the top hospitals in the world. It is one of the best places for patient care and medical research. Because of this, there is a great deal of employee satisfaction and job security at Mayo.

The Mayo Clinic in Rochester has about 1,500 doctors and about 318,690 patients. Apart from this, the Mayo Clinic treats about 1.38 million out-patients everyday. It has a staff of about 26,209 employees. Because of this large size, Mayo has a large number of departments that handle various aspects of the health care system.

1.2 Division of Medical Informatics Research

The Mayo Clinic is primarily a health service provider. This is where people go if they were really sick. *So, where do computer science graduates, like us, fit into a place like this?* This section describes the application of computer science to the field of medicine – Medical Informatics – and the department at which the internship was carried out.

The Mayo Clinic promotes the research in various fields related to medicine. In order to do this, it has set up the *Department of Health Sciences Research* (HSR). This department carries out research in a number of areas such as bio-statistics, bio-informatics, medical informatics, etc.

Medical Informatics is a field of medicine that deals with the effective storage and handling of medical data such as patient records, prescriptions, etc. It also includes the use of this data for knowledge discovery. Since this data is usually stored in plain text, Natural Language Processing techniques are applied to understand and make some sense of the data.

For example, doctors see number of patients everyday and either write or dictate the diagnoses of all these patients. Millions of such diagnoses are generated and stored everyday. It only makes sense to be able to use this huge data source to improve health care. It could be used, for instance, to determine possible diagnoses

and cures, given a set of unusual symptoms. It could also be used to prevent doctors from making mistakes made by doctors earlier, in treating certain symptoms.

The *Division of Medical Informatics Research* (MIR) is one of the divisions under the HSR. MIR consists of approximately 64 employees, including medical doctors, linguists, computer scientists and the like. The goals of this department are to provide "effective indexing and access to patient data". At the same time "understanding human disease through excellence in Biomedical Informatics" is an important task carried out by MIR.

1.3 History of Medical Informatics Research at Mayo

Research in Medical Informatics at Mayo dates back to 1907[1], when a unit for the storage and retrieval of medical records was created. A number of coding systems were used for the storage of medical records during that time. In 1935, the emerging technology at that time was embraced and the IBM punch card systems were employed to store medical records. The system was computerized in 1975. The department of Health Sciences Research was created in 1987, to initiate research in extracting knowledge from patient data.

The department has access to a rich data source, consisting of diagnoses and surgical procedures of all patients seen at the Mayo Clinic from 1909.

[1] Source: http://hsrwww.mayo.edu/medinf/history.html

2 The Project

Section 1 describes the Division of Medical Informatics Research at the Mayo Clinic. This department deals with research related to the efficient storage and retrieval of patient data and using this rich source of data to further the field of medicine, through knowledge discovery.

Since the department is dedicated to research, the project I worked on during my internship at the department was a research project. This section describes the project.

2.1 Previous Work

My Master's thesis work [6] in the Computer Science department at the University of Minnesota Duluth is in the broad field of Natural Language Processing. As part of this thesis I compared a number of measures of semantic relatedness with respect to human perception of relatedness and also with respect to their performance in a Natural Language Processing task. A new measure, based on context vectors was also introduced.

These measures used lexical resources such as WordNet and statistical information from large corpora of text to compute relatedness scores for pairs of concepts (For example, *paper–pencil*, *dog–cat*, etc). These measures included the following:

1. the Jiang-Conrath measure [3].
2. the Resnik measure [8].
3. the Lin measure [5].
4. the Hirst-St.Onge measure [2].
5. the Leacock-Chodorow measure [4].
6. the Extended Gloss Overlaps measure [1].
7. the Vector measure [6].

These measures were implemented as Perl modules [7] and distributed under the GPL on CPAN (an archive of Perl modules) as the WordNet::Similarity package. Each of these measures used WordNet as an electronic lexical resource for measuring semantic relatedness. Most of the measures combine the knowledge present in WordNet, with statistical data extracted from large corpora of text.

2.2 Medical Resources

A number of semantic networks and taxonomies of concepts exists in the medical informatics domain. For example, SNOMED CT® is a network of clinical concepts, connected by a number of relations. Similarly, MeSH is a taxonomy of medical Subject headings. These are very similar to the way WordNet is structured.

Other medical resources that are available at medical centers like the Mayo Clinic are large collections of patient diagnoses. These diagnoses are in the form of large collections of free flowing text.

The availability of such resources make it possible to use these in such tasks as measuring semantic relatedness of medical concepts.

2.3 Using a Medical Taxonomy to Measure Semantic Relatedness

In this internship, I modified the measures of semantic relatedness mentioned above, so as to use SNOMED CT® in place of WordNet. The measures could be then used to measure the semantic relatedness of medical concepts. However, rather than simply replacing the back-end taxonomy, the measures were generalized such that any semantic network of concepts or taxonomy can be used for measuring semantic relatedness. This enables us to compare the effectiveness of various ontologies for this task. It also facilitates the extension of these measures to other fields where such resources are available.

The measures were generalized by moving all WordNet-specific code and data into a separate module called the WordNet::Interface. This is an interface into WordNet. Similarly, interfaces for other semantic networks (such as SNOMED CT®) can be created, such that they conform to the rules laid out by the measures. All the semantic-network-specific data and code goes into these interfaces. The measures can then access these semantic networks through their corresponding interfaces and use these to measure the semantic relatedness of concepts present in those networks. Figure 1 show the schematic of the modules and how they interact

with one another.

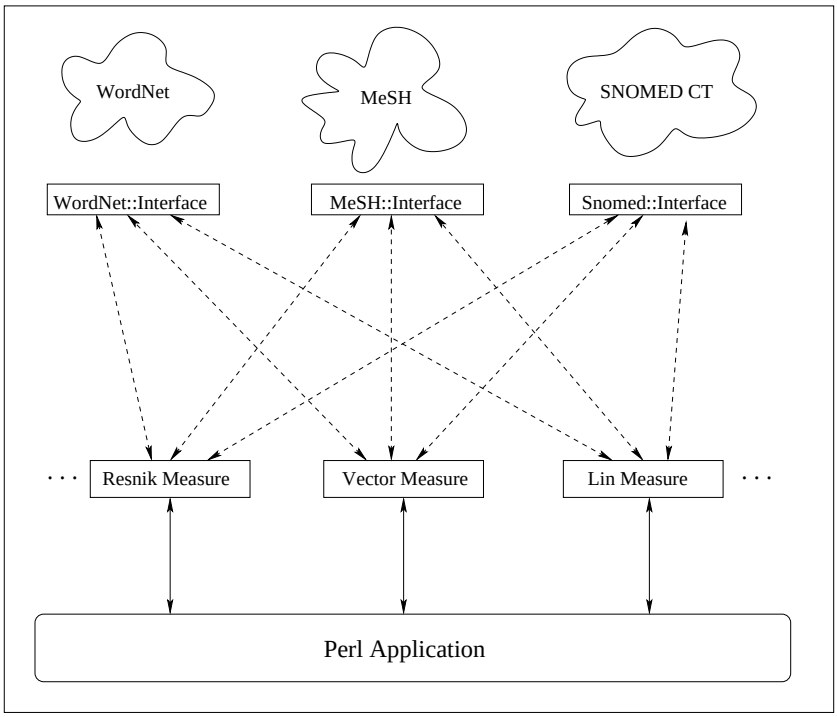

Figure 1: A schematic showing the interaction between the various modules.

At the Mayo Clinic, we first propose to measure the effectiveness of these measures in the Medical Informatics domain. If the measures perform as expected, they could then be applied to task such as semantically enriched information retrieval (of patient data), clustering of documents, building ontologies and the like.

3 Tools and the Environment

This section of the report basically describes the work environment – the hardware and the software used and the degree of supervision and training received during the internship.

3.1 Hardware and Software

I was assigned a Windows NT workstation for my work. However, a number of Sun machines running the Solaris operating system were available if we wished to work in a Unix environment. A Beowolf cluster running linux was also accessible. The Hummingbird X-Window system was installed on the Windows machine so as to be able to connect to and access the Solaris and the Linux machines.

All the modules were written in Perl, run on linux. Perl applications were written to test the modules. The SNOMED CT taxonomy was stored in a mysql database and was accessible through a mysql server running on the linux machine. This was accessed by the Perl modules using a Perl database driver module (DBD::mysql and DBI modules). The interface module to SNOMED CT (Snomed::Interface) used embedded SQL statements to access the data from the mysql server.

3.2 Work Environment

I was given very little training, since the work I was doing was very closely related to my thesis. I was closely supervised during the initial phase of my internship, while I was still learning the ropes. I received a lot of valuable advice from Dr. Serguei Pakhomov. For the rest of the internship I was pretty much on my own, with occasional meetings to get my supervisor up to speed with the work. At the end of the internship I gave a presentation of work (including a little demo of the software) to the entire department, which was followed by a discussion about the future prospects of the work.

4 Conclusions

This section highlights the important aspects of the internship.

1. The internship concluded on the 2nd of August 2003, and at the end of the two months I was able to produce a generalized version of the measures of semantic relatedness. However, due to the time constraints I was not able to document the software well enough.

2. The internship was a good learning experience. I was able to see the scope of research in the industry and the applicability of my field of research to real world problems.

3. I was able to get work experience in the industry, carry out research that interests me and earn some money all at the same time. Not to mention the credits I earn in school for this internship.

4. Mayo Clinic now has a useful tool that they can use in a number of tasks to improve the world of medicine.

5. I plan to continue the relationship with Mayo. For a start Dr. Pakhomov and I are planning to write a paper based on this work and submit it to the MedInfo 2004 conference in San Fransisco in September 2004.

6. My supervisor and colleagues in the department were quite impressed with the work and welcomed me to work with them in the future (perhaps next summer).

References

[1] S. Banerjee and T. Pedersen. Extended gloss overlaps as a measure of semantic relatedness. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence (IJCAI-03)*, Acapulco, Mexico, August 2003.

[2] G. Hirst and D. St. Onge. Lexical chains as representations of context for the detection and correction of malapropisms. In C. Fellbaum, editor, *WordNet: An electronic lexical database*, pages 305–332. MIT Press, 1998.

[3] J. Jiang and D. Conrath. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, Taiwan, 1997.

[4] C. Leacock and M. Chodorow. Combining local context and WordNet similarity for word sense identification. In C. Fellbaum, editor, *WordNet: An electronic lexical database*, pages 265–283. MIT Press, 1998.

[5] D. Lin. An information-theoretic definition of similarity. In *Proceedings of International Conference on Machine Learning*, Madison, Wisconsin, August 1998.

[6] S. Patwardhan. Incorporating dictionary and corpus information into a context vector measure of semantic relatedness. Master's thesis, Dept. of Computer Science, University of Minnesota, Duluth, 2003.

[7] S. Patwardhan and T. Pedersen. WordNet::Similarity modules version 0.05. Released, 2003. http://search.cpan.org/dist/WordNet-Similarity.

[8] P. Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal, August 1995.