Measuring the Similarity and Relatedness of Concepts in the Medical Domain : IHI 2012 Tutorial Overview

Ted Pedersen
Dept. of Computer Science
University of Minnesota
Duluth, MN 55812 USA
tpederse@d.umn.edu

Serguei Pakhomov
College of Pharmacy
University of Minnesota
Minneapolis, MN 55455 USA
pakh0002@umn.edu

Bridget McInnes
College of Pharmacy
University of Minnesota
Minneapolis, MN 55455 USA
bthomson@umn.edu

Ying Liu
College of Pharmacy
University of Minnesota
Minneapolis, MN 55455 USA
liux0395@umn.edu

ABSTRACT

The ability to quantify the degree to which concepts are similar or related to each other is a key component in many Natural Language Processing (NLP) and Artificial Intelligence (AI) applications. For example, in a document search application, it can be very useful to identify text snippets that contain terms that are similar to (but not identical) to those provided by a user. This tutorial will introduce the theory behind measures of semantic similarity and relatedness, and show how these can be applied in the medical domain by using freely–available open–source software[1] (UMLS::Similarity). This software takes advantage of the Unified Medical Language System[2] (UMLS), which is maintained by the National Library of Medicine (USA). The tutorial will also show how to evaluate existing measures with manually–created reference standards.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Natural Language Processing; J.3 [**Life and Medical Sciences**]: Medical Information Systems

General Terms

Experimentation, Algorithms

1. TUTORIAL TOPICS

This two hour tutorial is divided into three main areas, where examples from the medical domain will be used throughout to illustrate key ideas.

- **Semantic Similarity and Relatedness (30 minutes)**

  Dr. Pedersen will conduct a theoretical overview of semantic similarity and relatedness. He will discuss some of the most commonly used measures and explain the distinctions between them.

- **Using UMLS::Similarity (60 minutes)**

  Dr. McInnes and Dr. Liu will introduce the Unified Medical Language System and UMLS::Similarity [1], a freely–available open–source software package that computes measures of similarity and relatedness.

- **Getting Started (30 minutes)**

  Dr. Pakhomov will present a methodology for evaluating measures of similarity and relatedness using manually–created reference standards, and using the measures for clinical applications.

2. INTENDED AUDIENCE AND AIMS

This tutorial is designed for Health Informatics professionals with an interest in ontologies or Natural Language Processing. Those who attend will learn how to:

- understand the distinction between semantic relatedness and semantic similarity,
- measure semantic similarity and relatedness using information from ontologies, definitions, and corpora,
- access these measures from the UMLS::Similarity command line, API, and web services,
- integrate these measures into NLP / AI applications, and
- conduct experiments using freely–available manually–created reference standards.

3. ACKNOWLEDGMENTS

The development of this tutorial and the work that underlies it was supported in part by grant 1R01LM009623-01A2 from the National Library of Medicine, National Institutes of Health. The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

4. REFERENCES

[1] B. McInnes, T. Pedersen, and S. Pakhomov. UMLS-Interface and UMLS-Similarity : Open source software for measuring paths and semantic similarity. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 431–435, San Francisco, 2009.

[1] http://umls-similarity.sourceforge.net
[2] http://www.nlm.nih.gov/research/umls/

IHI'12, January 28–30, 2012, Miami, Florida, USA.
ACM 978-1-4503-0781-9/12/01.