A Cross-Domain Application of Natural Language Processing In Biology

Ivey Chiu
Department of Mechanical and Industrial Engineering
University of Toronto
Toronto, M5S 3G8, Canada
chiu@mie.utoronto.ca

Lily H. Shu
Department of Mechanical and Industrial Engineering
University of Toronto
Toronto, M5S 3G8, Canada
shu@mie.utoronto.ca

Abstract

Biomimetic design relies on relevant biological phenomena to serve as a basis for concepts in the engineering domain. Many instances of biomimetic design have resulted from personal observations of biological phenomena. However, a non-biologist's biology knowledge may be limited. To overcome this limitation, we perform keyword searches on existing natural-language knowledge sources. However, differences in domain lexicons present challenges to retrieving relevant information. A meaningful keyword to an engineer may not result in relevant matches within the biology. A method to systematically bridge these disparate domains is discussed as well as implications for any other cross-domain search applications.

1 Introduction

Engineers and designers have successfully applied concepts and principles found in biology to solve problems found in their own domains. Biomimetic design relies on the ability to draw analogies between the biological and engineering domains to generate engineering solutions. A well-known example is the invention of the Velcro fastener that was invented after observing that cockleburs attach to clothing and animal fur by small hooks. More recent biologically-based work includes correlation of heat transfer principles to shapes found in nature for optimization (Bejan, 2000). This appears to indicate that biology is a good source of inspiration for designers and engineers. However, it is not always feasible to rely on observation and personal knowledge of biological phenomena, as a non-biologist's biology knowledge may be limited. In an effort to systematically identify biological analogies for any given engineering problem, we have turned to existing biological knowledge sources in unstructured natural-language format.

Other current approaches to biomimetic design include compiling a database of biological phenomena indexed by engineering functions (Vincent & Mann, 2002; Lindemann & Gramann, 2004). This is a large task and can be subjected to the compiler's biases. These limitations, along with explosive information growth and database incompleteness (Rebholz-Schuhmann et al., 2005) are indicators that a non-database approach is required in our application as well as in bioinformatics.

Natural language analysis has been used to support and analyze design across many disciplines including architecture (de Vries et al. 2004), software engineering (Burg, 1997), and mechanical design (Yang & Cutkosky, 1997). These applications attempt to reduce design fixation, collect requirements and capture design information respectively.

2 Previous Work

The approach that has been developed by our laboratory is to perform a keyword search on a biology text (Vakili & Shu, 2001; Hacco & Shu, 2002) and to examine the matches retrieved for relevant biological phenomena. The initial text is *Life*, (Purves et al., 2001), a biology textbook used in a first-year biology course at the University of Toronto. The keywords used are verbs, as they convey functionality (Stone & Wood, 1999; Ullman, 2003), are not form specific and are central to the meaning of a sentence (Joanis & Stevenson, 2003).

3 Motivation

We encountered many of the challenges identified in computational linguistics such as word sense disambiguation (Yarowsky, 1995), anaphora resolution (Mitkov, 2001) and named entity identification (Li et al., 2004). Moreover, biologists and engineers have different domain-specific lexicons for describing their work. Therefore, a meaningful functional keyword for an engineer may result in few or no matches within the biology domain.

In a case study for "cleaning", e.g., removing dirt from clothing, a biochemist (Waygood, 2003) suggested "defend" as a possible functional keyword. He explained that many organisms clean or remove as a defensive mechanism. In a study, engineering students were presented with matches retrieved by the keyword "defend" and asked to generate concepts relating to cleaning clothes. Many of these students produced successful concepts (Mak & Shu, 2004), including modular clothing where the dirty parts are removed. An excerpt from *Life* (Purves et al., 2001) located by searching for all forms of "defend" follows:

> When pathogens pass these barriers, plant **defenses** are activated. Plants seal off and sacrifice the damaged tissue so that the rest of the plant does not become infected. This approach works because most plants can replace damaged parts by growing new stems, leaves and roots.

While "defend" produced relevant phenomena, it was unclear at this point how to automatically generate such a biologically meaningful word without expert assistance. Use of WordNet (2.0) and a thesaurus (Manser, 2004) did not produce a direct route between "clean/remove" and "defend". The challenge is to systematically bridge the differences between engineering and biology lexicons.

4 Studies and Results

In the following, we describe how we increase the search space and how words retrieved relate to the original keyword. From this, we found a set of different verbs that appear to bridge the two domains and examined methods to manage these verbs. A part-of-speech tagger (Brill, 1994) and a partial parser (Abney, 2002) were used for this work.

4.1 Generating alternative keywords

We first focused on expanding the search space by generating other functional keywords using WordNet (2.0) as a language framework. We chose to use troponyms as they describe a specific manner of accomplishing tasks (Fellbaum, 1993). For example, "sauntering" is a specific manner of "walking", and "cleaning" is a specific manner of "removing" (WordNet, 2.0). Using troponyms rather than synonyms of the keyword enabled us to generate alternative keywords that improved the quantity and quality of matches. It also provided some keywords not obviously related to the initial keyword of "clean" or "remove", e.g., "excrete", "eliminate", "kill" and "draw" as in "to draw water" through capillary action (WordNet, 2.0). *Life* was searched with all troponyms of "remove".

The contents of the matches retrieved by the keywords were analyzed to find relevant biological phenomena using concepts of frequency distribution (Zipf, 1949) and collocation (Yarowsky, 1995; Banerjee & Pedersen, 2003). High frequency words from the matches were examined and often found to be agents or objects of the keyword. Together, the agents, objects and keywords describe biological concepts associated with that functional keyword. For example, high frequency words were "predator", "prey" and "species" for the keyword "eliminate", thus describing how interactions between prey and predator species lead to one another's elimination. Details can be found in Chiu & Shu (2004).

4.2 Bridging disparate domains

It was noticed that, while the high frequency words are often attached to the keyword as agents or direct objects, there were many instances where high frequency words are attached to different verbs.

As a result, we turned our attention to identifying the non-keyword verbs in these instances. For example, all matches retrieved by "kill" had frequent words "cells" and "body". From the excerpt below (Purves et al. 2001), it can be seen that "cells" is attached to "kill" and "body" is attached to "defend" as an object.

> As HIV **kills** more and more TH ***cells***, the immune system is less and less able to defend the ***body***…

The verbs identified and collected this way are called *bridge verbs*, as we believe this process provides a basis to bridge the biology and engineering domains.

4.3 Organizing and correlating the bridge verbs

We sought a method to organize the bridge verbs without discarding any. Various properties of the verbs were examined including word frequencies in written English (Leech, Rayson & Wilson, 2004), verb class (WordNet, 2.0), and *biological significance* as measured using the biology terms of two biological references (Hine & Martin, 2004, Biology-online, 2005). For example, the term "diffusion" identifies the bridge verb "diffuse" as biologically significant.

Based on these properties and lexical interrelationships documented in WordNet (2.0), word graphs were generated to organize the bridge verbs. Word graphs have also been used to support the architectural design process (de Vries et al. 2004). Word graphs show naturally forming clusters of words depicting interrelationships in a way not possible with flat lists. By following arcs representing relationships from word to word, we discovered alternative keywords.

We proceeded to correlate biological significance with word counts from dictionary definitions in Biology-online.org (2005) rather than terms, as it was noticed many seemingly meaningful words were not included in terms alone, such as "defend." Correlating the bridge verbs with dictionary count is based on the rationale that authors treat their subject matter with a minimal set of words to convey specific meanings (Zipf, 1949; Luhn, 1959). Their research suggests that word use follows a distribution such that it is possible to predict meaningful words based on frequency.

Correlating bridge verbs with dictionary counts showed that a small range of dictionary counts corresponded with the majority of biologically significant words. Consequently, it appears that many words within this range of dictionary counts are biologically meaningful, regardless of whether they were explicitly identified as biologically significant or not. Therefore, the words within the range provide designers with the most relevant biological keywords for the biomimetic search process. "Defend" as suggested by the biochemist is located within this range. A follow-up study indicates that this process can systematically generate biologically meaningful keywords.

5 Summary and concluding remarks

Biomimetic design relies on relevant biological phenomena to serve as a basis for engineering solutions. We use existing biology knowledge in an unstructured natural-language format to facilitate and support the biomimetic design process so that engineers need not rely on their own knowledge of biology. However, differences in domain lexicons challenge the ability to retrieve relevant biological phenomena. Using a combination of collocation and frequency analysis, we found a method to generate biology keywords related to the original engineering keyword. Different methods of organizing and correlating the set of bridge verbs were explored to provide the engineer with the most relevant words first.

This method of generating bridge verbs is generic and can be applied to systematically bridge any two disparate domains (e.g., engineering and economics). This enables problems from one domain to be solved using ideas from a different domain thus promoting creative problem solving and the generation of novel solutions.

Acknowledgement The authors gratefully acknowledge the financial support of NSERC (Natural Sciences and Engineering Research Council of Canada). The authors also acknowledge the generosity of Biology-online.org and Purves et al. for providing machine-readable documents.

References

Steven Abney. 2002. *SCOL version 1h.* www.vinartus.net/spa/.

Satanjeev Banerjee and Ted Pedersen. 2003. The Design, Implementation and Use of the Ngram Statistics

Package. *Proc. CICLING-03, 02/17-21, Mexico City, Mexico.*

Adrian Bejan. 2000. From Heat Transfer Principles to Shape and Structure in Nature: Constructal Theory. *Transactions of the ASME,* 122:430-449.

Biology-online. 2005. www.biology-online.org.

Eric Brill. 1994. *Rule-based part-of-speech tagger*. www.cs.jhu.edu/~brill/.

J.F.M. Burg. 1997. *Linguistic Instruments In Requirements Engineering*. Vrije Universiteit, Amsterdam, The Netherlands, Ph.D. thesis.

Ivey Chiu and Lily H. Shu. 2004. Natural Language Analysis for Biomimetic Design. *Proc. ASME DETC/CIE,* Salt Lake City, UT, DETC2004-57250.

Bauke de Vries, Joran Jessurun, Nicole Segers and Henri Achten. 2004. Word Graphs in Architectural Design. *Proc. International Conference on Design Computing and Cognition 2004*, pp 541-556.

Christiane Fellbaum. 1993. English Verbs as a Semantic Net. *Five Papers on WordNet,* pp 40-61. ftp://ftp.cogsci.princeton.edu/pub/wordnet/5papers.ps

Eli Hacco and Lily H. Shu. 2002. Biomimetic Concept Generation Applied to Design for Remanufacture. *Proc. ASME DETC/CIE,* Montreal, QC, Canada, DETC2002/DFM-34177.

Robert S. Hine and Elizabeth Martin (eds). 2004. *A Dictionary of Biology*. Oxford University Press.

Eric Joanis and Suzanne Stevenson. 2003. A General Feature Space for Automatic Verb Classification. *Proc. 10th Conference of the European ACL,* pp. 163-170.

Geoffrey Leech, Paul Rayson and Andrew Wilson. 2001. *Companion Website for: Word Frequencies in Written and Spoken English: based on British Nat. Corpus*. www.comp.lancs.ac.uk/ucrel/bncfreq/.

Xin Li, Paul Morie and Dan Roth. 2004. Identification and tracing of ambiguous names: Discriminative and generative approaches. *Proc. National Conference on Artificial Intelligence*. pp. 419-424, San Jose, CA.

U. Lindemann and J. Gramann. 2004. Engineering Design Using Biological Principles. *Proc. International Design Conf. – Design 2004,* Dubrovnik 5/18-21.

Hans P. Luhn. 1959. Auto-Encoding of Documents for Information Retrieval Systems. *Modern Trends in Documentation,* Pergamon Press, London, pp. 45-58.

Teresa W. Mak and Lily H. Shu. 2004. Use of Biological Phenomena in Design By Analogy. *Proc. ASME DETC/CIE,* Salt Lake City, UT, DETC200-57303.

Martin H. Manser (ed). 2004. *The Chambers Thesaurus*. Chambers Harrap Publishers Ltd, Edinburgh, UK.

Ruslan Mitkov. 2001. Towards a more consistent and comprehensive evaluation of anaphora resolution algorithms and systems. *Applied Artificial Intelligence,* 28(3):253-276.

William K. Purves, David Sadava, Gordon H. Orians, H. Craig Heller. 2001. *Life, The Science of Biology*, 6/e, Sinauer Associates, Sunderland, MA.

Dietrich Rebholz-Schuhmann, Harald Kirsch and Francisco Couto. 2005. Facts from Text – Is Text Mining Ready to Deliver? *PLoS Biology*, 65:188-191

Robert B. Stone and Kristin L. Wood. 1999. Development of a Functional Basis for Design. *Proc. ASME/DETC/CIE*, Las Vegas, NV, DETC99/DTM-8765

David G. Ullman. 2003. *The Mechanical Design Process, Third Edition*, McGraw-Hill, New York, NY.

Vanessa Vakili and Lily H. Shu. 2001. Towards Biomimetic Concept Generation. *Proc. ASME/CIE*, Pittsburg, PA. DETC2001/DTM-21715

Julian Vincent and Darrell Mann. 2002. Systematic Technology Transfer from Biology to Engineering, *Philosophical Transactions of The Royal Society: Physical Sciences*, 360:159-173

E. Bruce Waygood. 2003. Coordinator of Health Research, University of Saskatchewan, Personal communication.

WordNet. 2.0. http://www.cogsci.princeton.edu/~wn/.

Maria C. Yang and Mark R. Cutkosky. 1997. Automated Indexing of Design Concepts for Information Management. *Proc. International Conference on Engineering Design*, Tampere, Finland 08/19-21.

David Yarowsky. 1995. Unsupervised word-sense disambiguation rivalling supervised methods. *Proc. of 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189-196.

George K. Zipf. 1949. *Human Behavior And The Principle of Least Effort; An Introduction To Human Ecology*, Addison-Wesley Press, Cambridge, Ma.