Proceedings of the 14th International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/CIE 2012
August 12-15, 2012, Chicago, Illinois, USA

DETC2012-70732

AUTOMATIC EXTRACTION OF CAUSALLY RELATED FUNCTIONS FROM NATURAL-LANGUAGE TEXT FOR BIOMIMETIC DESIGN

Hyunmin Cheong
cheong@mie.utoronto.ca

L. H. Shu*
shu@mie.utoronto.ca
*Corresponding author

Dept. of Mechanical & Industrial Engineering
University of Toronto
5 King's College Road
Toronto, ON, M5S 3G8, Canada

ABSTRACT

Identifying relevant analogies from biology is a significant challenge in biomimetic design. Our natural-language approach addresses this challenge by developing techniques to search biological information in natural-language format, such as books or papers. This paper presents the application of natural-language processing techniques, such as part-of-speech tags, typed-dependency parsing, and syntactic patterns, to automatically extract and categorize causally related functions from text with biological information. Causally related functions, which specify how one action is enabled by another action, are considered important for both knowledge representation used to model biological information and analogical transfer of biological information performed by designers. An extraction algorithm was developed and scored F-measures of 0.78-0.85 in an initial development test. Because this research approach uses inexpensive and domain-independent techniques, the extraction algorithm has the potential to automatically identify patterns of causally related functions from a large amount of text that contains either biological or design information.

1. INTRODUCTION

Biomimetic design uses biological phenomena as inspiration for solutions to engineering problems. Although many innovative biologically inspired design solutions have been developed, the sources of inspiration are often limited by the designer's knowledge in biology or chance observation. To address this challenge, the research group led by Shu (2010) has used a natural-language approach to identify relevant biological analogies for more than a decade.

The natural-language approach takes advantage of the enormous amount of information already available in text format. Strategies and techniques are developed to search, identify, and categorize relevant biological information from books, papers, and online resources, etc. The approach allows designers to identify useful biological analogies beyond those indexed specifically for engineering design.

On the other hand, the amount of available information poses a challenge as well. Even from a single corpus, certain keywords can retrieve a large number of search results that designers may find overwhelming and irrelevant. In addition, Mak and Shu (2008) and Cheong and Shu (2009) have observed that designers often experience difficulties in identifying and transferring analogies from text descriptions of biological phenomena.

To further support the natural-language approach, this research aims to automatically extract causally related functions from natural-language text. Cheong and Shu (2009) noted that causally related functions in the descriptions of biological phenomena, which specify how one action is enabled by another action, may help designers identify and apply the relevant analogies. In general analogical reasoning, Gentner (2006) states that finding similarites of higher-order relations, such as causal relations, plays a key role in successful analogical transfer. We used computational linguistic techniques and a set of syntactic patterns to extract causally related functions from a biological corpus. The extracted information could then be categorized by the enabling functions of causal relations.

Section 2 highlights why extracting causally related functions could be relevant not just for biomimetic design, but also for other goals in engineering design. Section 3 describes

the method developed to extract causally related functions and Section 4 reports the results from development testing. Section 5 discusses the potential benefits of causal-relation extraction for concept generation and other design studies. Lastly, challenges and conclusions observed from this research are presented.

2. BACKGROUND

This section highlights how causal relations play an important role in both the modeling and natural-language approaches to biomimetic design. Next described is the application of computational linguistics in causal-relation extraction and other design research.

2.1. Modeling approach to biomimetic design

A number of formal representations have been used to model biological information for biomimetic design. Goel et al. (2009) used the structure-behavior-function framework to capture causal processes between states of biological systems. Chakrabarti et al. (2005) developed a more specific model of causality based on SAPPHiRE constructs (parts, state, organ, physical effect, input, physical phenomenon, action). Nagel et al. (2010) used the functional basis terms to model biological systems and index them in a design repository.

It should be noted that both Goel et al. and Chakrabarti et al. attempted to represent causality in biological phenomena. The formal representation of causal relations helps designers transfer the complex information of biological systems to engineering solutions. However, the use of these tools is limited by the amount of biological information that must be indexed. In particular, identifying the candidate biological information to be indexed is a challenging task.

2.2. Natural-language approach to biomimetic design

To identify useful biological analogies for biomimetic design, our research group has focused on searching biological information described in natural-language format. Various computational linguistic techniques have been applied to support our work.

2.2.1. Searching natural-language text

Hacco and Shu (2002) suggested using a Brill tagger and WordNet to distinguish relevant search results from irrelevant ones. Chiu and Shu (2007) used word frequency, collocation, and WordNet to develop a procedure that identifies candidate biologically meaningful keywords. Cheong et al. (2011) then extended the procedure to translate the functional terms of the functional basis (Stone and Wood 2000) to biologically meaningful keywords. Biologically meaningful keywords are terms that are deemed to be more useful in searching biological text for analogies than the corresponding engineering keywords.

2.2.2. Benefits of the natural-language approach

Shu et al. (2011) present several application case studies of the natural-language approach. The studies demonstrated that the natural-language approach could identify nonobvious analogies based on the transfer of abstract strategies. Many existing examples of biologically inspired design only involve the direct mimicry of biological systems, which includes similarity transfer at both geometric and strategic levels, e.g., Velcro, gecko feet, legged robots, etc. Our application case studies, on the other hand, applied the strategy of abscission in the assembly of microparts or the strategy of preemptive failure in a sacrificial snap-fit design. We believe that designers are more likely to find these nonobvious analogies when they use functional keywords to locate analogies from a large number of sources.

2.2.3. Limitations of the natural-language approach

Limitations of the natural-language approach include the need to process a potentially large number of search results. To address this challenge, Ke et al. (2010) applied part-of-speech tags and word sense disambiguation based on the WordNet taxonomy to reduce the number of relevant search results. While these techniques eliminated some irrelevant search results, designers still need to identify which matches include useful analogies to their problems.

Vandevenne et al. (2011) also point out that the translation of biologically meaningful keywords is not fully scalable. Because a fair amount of manual processing is still necessary in the translation process, updating or generating a new list of keywords would be resource intensive. Vandevenne et al. suggest using a more scalable approach based on the analysis of term occurrences in biological text.

2.2.4. Causal relations in the natural-language text

This research uses automatic extraction and categorization of causal relations to address the limitations mentioned above.

Cheong et al. (2011) observed that a set of semantic relations usually hold between biologically meaningful keywords and the corresponding engineering keywords in biological text. One of these semantic relations involved causally related functions. In the sentence “Lysozymes destroy bacteria to protect animals,” the verbs “destroy” and “protect” are causally related. If the designer originally wanted to find out how “protection” occurs in biology, the causal relation identifies that “destroying” enables “protection.” Also, the relation provides a structural framework that facilitates the transfer of relevant functions from the biological phenomenon (Gentner 1983).

In an empirical study, Cheong and Shu (2009) observed that text descriptions of biological phenomena containing causally related functions are more likely to serve as useful analogies for design problems. Based on these findings, we hypothesized that extracting and categorizing causally related functions in natural-language search results could help designers identify relevant biological analogies. To automate the extraction process, we applied natural-language processing techniques developed in computational linguistics.

2.3. Causal-relation extraction in computational linguistics

Computational linguistic researchers have tried to automatically extract semantic information from English text with varying degrees of success. For extracting causal relations, two different approaches have been used.

Joskowiscz et al. (1989) and Kaplan and Berry-Rogghe (1991) manually coded sentences into machine-readable propositions. Computer algorithms would then identify causality based on the propositions entered. The main limitation of these approaches is that domain-specific hand-coding makes scaling up for other applications difficult.

Recent approaches have used linguistic patterns to identify explicit causal relations in text. Garcia's (1997) automatic algorithm looks for a set of causative verbs, e.g., "causes," as cues to identify causality in text. In addition to causative verbs, Khoo et al. (2000) used causal links, e.g., "because" or "therefore," as the cues for causal relations. Khoo et al. also used a parse tree to identify particular syntactic patterns that likely represent causal relations. Similarly, Girju (2003) used causal verbs and a set of lexico-syntactic patterns to identify cause and effect from text with the goal of developing a question-answering system.

Our research also uses linguistic patterns to identify causal relations in text. However, our objective differs from the previous research in computational linguistics because we want to extract causally related functions, which are more relevant to solving design problems. The previous researchers used a set of *explicit* linguistic cues, e.g., causative verbs, to determine causal relations between concepts. Our work attempts to only use syntactic patterns to identify *implicit* causal relations between verbs.

3. EXTRACTING CAUSALLY RELATED FUNCTIONS

Cheong et al.'s (2011) previous work on identifying biologically meaningful keywords provides a framework for defining linguistic patterns that represent causally related functions in biological text.

3.1. Computational linguistic tools used

In order to automatically identify relevant linguistic patterns, Stanford part-of-speech tagger v3.0 by Toutanova and Manning (2000) and Stanford parser v1.6.7 by de Marneffe et al. (2006) were used to tag and parse the text of interest. The tagger analyzes an English sentence and identifies a Penn Treebank part-of-speech tag (Marcus et al. 1993) for each word in the sentence. The parser then identifies grammatical relations, also known as typed dependencies, between a pair of words in the sentence. The tagger is about 97% accurate with trained text, while the parser has recorded an accuracy in the high 80's. Table 1 shows a sentence analyzed by the tagger and the parser.

3.2. Syntactic patterns of causally related functions

Life by Purves et al. (2001), a reference text for an entry-level university biology course, was chosen as the corpus. Seven chapters were manually read to code causally related functions found. Each chapter was chosen from a different section of the corpus (there are seven sections), to examine sections of text that describe a variety of topics in the corpus. Based on this coding process, six syntactic patterns that contain causally related functions were identified (Table 2).

Table 1: Tagging and parsing of an example sentence.

Example sentence:	
"Lysozymes destroy bacteria to protect animals."	
Tagged:	
Lysozymes/NNS destroy/VBZ bacteria/NNS to/TO protect/VB animals/NNS ./.	
NNS: noun, plural	*TO: preposition "to"*
VB: verb, base form	
VBZ: verb, 3rd person, singular, present	
Parsed, typed dependencies:	
nsubj(destroy-2, Lysozymes-1)	
nsubj(protect-5, bacteria-3)	aux(protect-5, to-4)
xcomp(destroy-2, protest-5)	dobj(protect-5, animals-6)
nsubj: normal subject	*dobj: direct object*
xcomp: open clausal complement	*aux: auxiliary*

3.3. Automatic extraction and categorization of causally related functions

3.3.1. Processing the corpus

The entire corpus was first pre-processed to replace special characters that can cause parsing or tagging errors. The pre-processing algorithm also inserted a period at each line break without a period, e.g., chapter/section titles, because the parser looks for periods to determine the end of a sentence or phrase. In addition, we had to manually replace dashes that are used to set off a word/phrase with commas, because the parser identifies dashes as hyphens and treats these instances as compound words. The replacement of dashes was the only manually intensive processing task.

The pre-processed corpus was then automatically tagged and parsed with the Stanford tagger and parser.

3.3.2. Writing the extraction algorithm

A causal-relation extraction algorithm was written in Perl to read the part-of-speech tags and dependency relations and identify the six syntactic patterns in Table 2. The algorithm mainly looks for the types of dependency relations and uses the part-of-speech tags to test additional rules. For example, for Pattern #1 in Table 2, the algorithm checks to see if there is a word tagged as NN (noun) between the verbs identified in the dependency relation. For Patterns #4 and #6, the algorithm checks if the first verb **"**destroy**"** is tagged as VBG (gerund/present participle). In addition, if the enabling verb was initially identified as one of the auxiliary verbs in English, e.g., "need to," "have to," etc., the algorithm identifies the main verb that follows the auxiliary verb, e.g., "need to **protect**," as the enabling verb.

Table 2: Syntactic patterns of causally related functions found in *Life* (Purves et al. 2001). "DR" stands for "dependency relation."

1. Lysozymes **destroy** bacteria to **protect** animals.
   - The verb "protect" is an open clausal complement (DR: **xcomp)** to the verb "destroy." In other words, "protect" does not have its own subject, but has the same subject as "destroy."

     Exception: When the first verb is intransitive, i.e., does not have an object, the verbs are usually not causally related. For example, "I like to swim" does not express any causality although "swim" is defined as an open clausal complement to "like."

2. Bacteria are **destroyed** to **protect** animals.
   - Similar to Pattern #1, the verb "protect" is an open clausal complement (DR: **xcomp**) to the verb "destroy." In this case however, the main verb "destroy" is in the passive voice and the exception rule for Pattern #1 is ignored.

3. Lysozymes **destroy** bacteria, **protecting** animals.
   - The verb "protect" is an open clausal complement (DR: **xcomp**) to the verb "destroy."

4. By **destroying** invading bacteria, lysozyme **protects** animals.
   - The gerund "destroying" is a prepositional clausal modifier of the verb "protect," linked with the preposition "by" (DR: **prepc_by**).

5. To **protect** animals, lysozymes **destroy** bacteria.
   - The verb "protect" is part of a purpose clause modifier "To protect animals," which specifies the purpose of the following clause "lysozymes destroy bacteria" (DR: **purpcl**).

6. **Destroying** bacteria **protects** animals.
   - The gerund "destroying" acts as a clausal subject for the verb "protect" (DR: **csubj**).

3.4. Filtering out matches with non-physical descriptions

Hacco and Shu (2002) first reported that matches with keyword verbs acting on abstract objects, e.g., "support the *theory*," are less useful for biomimetic design. Some causal relations identified had enabling or desired functions acting on abstract objects, as in the following sentence from Purves et al. (2001):

"Scientists have now **found** morphological *evidence* to **support** the *theory*."

Ke et al. (2010) used the WordNet taxonomy to determine if a particular noun is abstract or physical. At the highest level of the taxonomy, all nouns are classified as either an "abstract entity" or a "physical entity." We also used this WordNet classification to distinguish whether nouns are abstract or physical. All the nouns in the corpus, which were identified with the tagger, were compared against the WordNet (3.0) noun hierarchy. If all the senses of a noun were classified under "abstract entity," the noun was included in a noun stop list. If any noun from the stop list appeared as the object of enabling or desired functions, those results would be filtered out. Creating the noun stop list is automatic and only needs to be performed once each time a new corpus is introduced.

3.5. Removing matches with non-meaningful verbs

The extraction algorithm retrieved some matches with causally related verbs that may not be meaningful for biological analogies. This section describes the method used to remove the non-meaningful verbs.

3.5.1. Light verbs

Manning and Schutze (1999) define a light verb as one that has little semantic meaning of its own, but becomes more meaningful when combined with an object, e.g., "take" vs. "take a walk." A particular light verb that was most frequently found in the corpus is "use." The following sentence from Purves et al. (2001) contains "use" as its enabling function:

"An enzyme **uses** energy from ATP hydrolysis to **unwind** the DNA."

Although the verb "uses" itself has little semantic context of its own, the phrase "uses energy" describes a meaningful process that enables the function of "unwinding." A challenge arises when categorizing causal relations that contain "use" as the enabling function. Because the light verb "use" takes different meanings depending on the associated object, matches with "use" as the enabling function can contain biological phenomena with varying contexts, e.g., "use energy" vs. "use proteins." For future research, we plan to investigate categorizing matches with light verbs used as enabling functions. For now, we exclude these matches to focus on causal relations between functionally or biologically meaningful verbs.

3.5.2. Simple causative verbs

Girju (2003) reported that a set of simple causative verbs, e.g., "cause," "lead to," "allow," etc., could be used to identify explicit causality in text. These causative verbs, similar to the light verbs, can convey different meanings in a sentence depending on the associated subject and object. Again for now, as we are interested in causal relations formed between meaningful functions, we exclude matches that contain simple causative verbs as enabling functions. In the future, the search algorithm could include rules to identify and categorize explicit causal relations defined by causative verbs.

3.5.3 Frequently appearing verbs

Manning and Schutze (1999) suggest that the most frequently appearing words in a corpus are likely to be

semantically weak, i.e., not meaningful. The most frequent verbs in our corpus include "use," "need," "enable," and "allow." We noticed that most of these verbs are either light verbs (e.g., "use" or "need") or simple causative verbs (e.g., "enable" or "allow"). Other frequent verbs were those used in describing scientific studies, e.g., "perform," "find," or "study," etc. Based on this observation, we decided to create a verb stop list that identifies and removes matches containing semantically weak verbs based on their frequency in the corpus.

3.5.4. Creating a verb stop list

Creating a stop list only based on verb frequency was not ideal because frequent verbs of the corpus included biologically meaningful keywords such as "bind" or "release." In order to identify a more pertinent stop list, the following procedure was applied:

1) A chapter that mostly describes scientific studies, rather than physical phenomena of biology, was "Chapter 21: History of Evolution." Manual reading of the chapter confirmed that most causal relations in the chapter are not meaningful.

2) Verbs that are associated in the syntactic patterns of causal relations (Table 2) are identified from Chapter 21. Because most causal relations of Chapter 21 were deemed to be not meaningful in Step 1, the verbs associated with these causal relations are likely to be semantically weak.

3) To identify semantically weak verbs for the entire corpus, the verbs identified from Chapter 21 in Step 2 were used as keywords to extract causal relations from the entire corpus. The enabling functions of these causal relations were then gathered. If a causal relation contained a semantically weak verb as its desired function, the associated enabling function was observed to be not meaningful in most cases. For example, in the sentence "Ecologists **study** patterns of distribution of organisms to **find** out how they change over time," both "study" and "find" are considered semantically weak.

4) The final verb stop list was constructed. The list includes semantically weak verbs that are identified at least two times in Step 3. The verbs identified only once in Step 3 were considered to be in the lower cut-off range of the Zipf's rank-frequency distribution (Manning and Schutze 1999). Such verbs on the lower cut-off range may be biologically specific and meaningful verbs.

Table 3 shows the semantically weak verbs identified. With the stop list implemented, the number of search results decreased significantly. For the search keyword "move," the stop list filtered out 41 of 74 results, potentially missing some relevant information. However, because one of the goals was to reduce a large number of natural-language search matches for designers, we decided to first develop an algorithm with higher precision than recall.

Table 3: List of semantically weak verbs identified.

able	do	lead	seek
act	enable	live	seem
allow	expect	make	serve
appear	express	manipulate	set
be	fail	modify	study
associate	find	need	take
attempt	function	perform	tend
begin	gather	permit	test
believe	go	present	think
call	happen	proceed	try
cause	have	process	undergo
coax	help	produce	use
come	improve	require	want
continue	include	rise	wish
depend	interact	say	work
deprive	learn		

3.6. Implementing the algorithm in a search tool

A search tool was developed based on the causal-relation extraction algorithm. The tool takes a verb keyword as input and retrieves causal relations from *Life* (Purves et al. 2001) where the keyword is the desired function. The results are then categorized by enabling functions of causal relations retrieved, and displayed based on the frequency of enabling functions. Figure 1 shows results of an example search.

3.6.1 Subject/object categorization

In addition to causal-relation extraction, the search tool categorizes the results of verb searches by the verb's subject or object. The search algorithm identifies subjects and objects of verb keywords with dependency relations. Figure 2 shows an example of object categorization. With subject/object categorization, designers could more easily identify relevant search matches for their design problems. For example, the "prevent" + water grouping in Figure 2 would be especially useful for developing waterproof design solutions.

3.6.2. Adjective search

The search tool also supports adjective searches and categorizes search results by the nouns modified by the adjective. Ke et al. (2010) used adjectives as keywords to search for biological analogies because adjectives describe qualities of problems or possible solutions. If designers were seeking a solution that must function in a dry environment, search results obtained with the adjective keyword "dry" could be useful. Figure 3 shows search results of the adjective keyword "dry," categorized by the nouns that the keyword is modifying. The search algorithm identifies the modified nouns of adjectives**.**

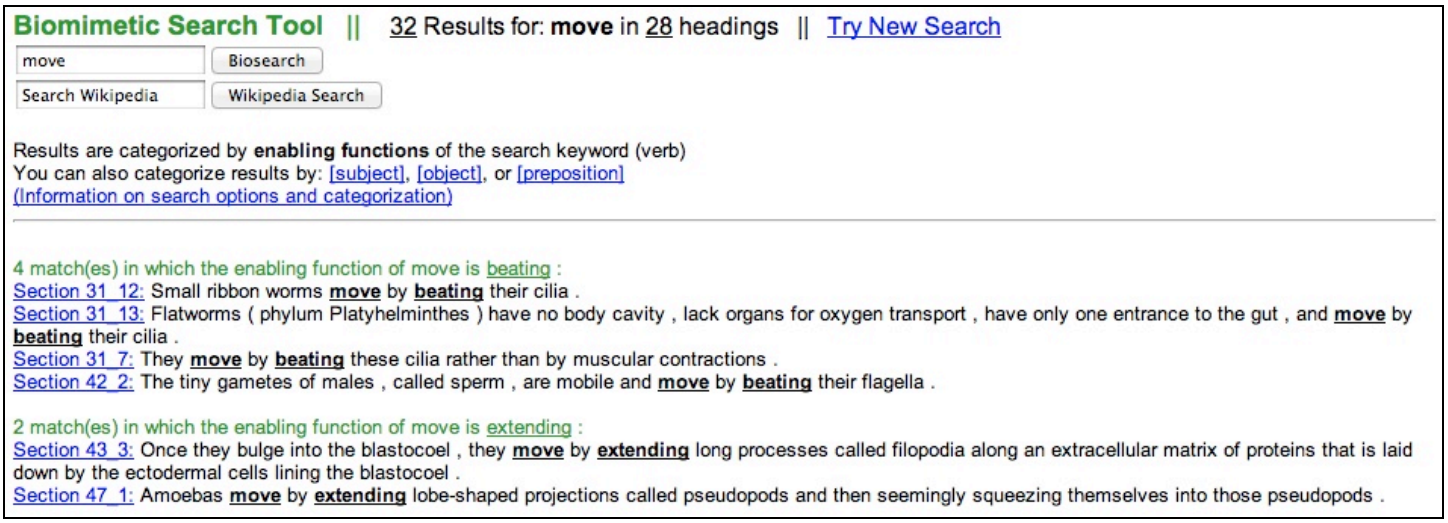

Figure 1: Search results for verb keyword "move," categorized by the enabling functions of "move."

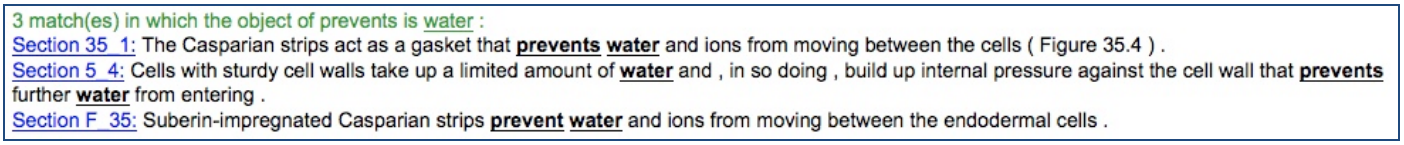

Figure 2: Search results for verb keyword "prevent," categorized by the object "water."

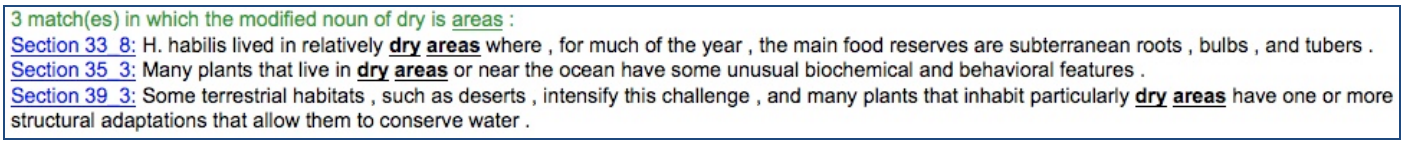

Figure 3: Search results for adjective keyword "dry," categorized by the modified noun "areas."

4. DEVELOPMENT TEST RESULTS

The goal of development testing was to examine the accuracy of the extraction algorithm and identify sources of error that must be removed.

4.1. Test method

Three chapters were randomly chosen for development testing. The lead author manually read each chapter to mark relevant causally related functions within a single sentence. Causally related functions were considered to be relevant if: 1) the enabling function was one of the biologically meaningful keywords identified from the functional basis translation work (Cheong et al. 2011) and 2) the causal relation was based on one of the syntactic patterns in Table 2. Using this set of relevant causal relations identified, precision, recall, and F-measures of the extraction algorithm were calculated. Precision and recall are defined below:

$$\text{Precision} = \frac{(\#\text{ of correctly retrieved causal relations})}{(\#\text{ of causal relations retrieved})}$$

$$\text{Recall} = \frac{(\#\text{ of correctly retrieved causal relations})}{(\#\text{ of causal relations in text})}$$

F-measure considers both precision and recall to compute the accuracy of a retrieval algorithm:

$$F = \frac{2*(\text{Precision})*(\text{Recall})}{(\text{Precision})+(\text{Recall})}$$

4.2. Comparison against the baseline performance

Manning and Schutze (1999) define the baseline measure as the performance of the simplest possible algorithm. The baseline measure indicates how difficult it is to improve a particular computational linguistic task. For our research, the baseline algorithm identifies co-occurring verbs in a sentence as causally related functions.

4.3. Overall test results

Table 4 presents test results for the extraction algorithm. For all three chapters, high precision and moderate recall scores were found. F-measures varied from 0.780 to 0.848, indicating promising accuracy for the extraction algorithm. The relatively narrow range of F-measures also suggests the consistency of the extraction algorithm over text on different topics in biology. The baseline algorithm by comparison had F-measures from 0.400 to 0.579, suggesting that the extraction task is not trivial.

Table 4: Comparison of precision, recall and F-measures between extraction and baseline algorithms in three chapters.

		Precision	Recall	F-measure
Ch. 4 (hits=21)	Extraction	0.913	0.778	0.840
	Baseline	0.423	0.815	0.557
Ch. 35 (hits=14)	Extraction	0.993	0.778	0.848
	Baseline	0.550	0.611	0.579
Ch. 44 (hits=16)	Extraction	0.800	0.762	0.780
	Baseline	0.308	0.572	0.400

Some relevant results were missed mostly because of parsing errors. Table 5 shows the sources of error for missed information. The accuracy of the algorithm is therefore limited by the accuracy of the parser. On the other hand, we found instances when causally related functions are not associated by a single dependency relation, but indirectly associated through multiple dependency relations. More study is required to develop algorithm rules that detect these complex functional relations.

Table 5: Sources of error for missed information.

Source of error	Freq.
Parser error: incorrectly identified the part-of-speech of a relevant verb	6
Parser error: could not find dependency relations between causally related verbs	3
Algorithm error: a complex causal relation not captured by the algorithm	3
Parser error: incorrectly identified the dependency relation between causally related verbs	2
Algorithm error: the verb stop list removes relevant causal relations	2

The limitation of this initial development test was that the patterns of causally related functions (Table 2) were created based on the lead author's analysis. Therefore, the precision and recall scores reported in this paper indicate the accuracy of the algorithm in capturing the intended information defined in Table 2. The final testing can incorporate multiple coders to independently identify causally related functions that they consider meaningful. Because determining which causal relations could be useful for design-by-analogy can be an ambiguous task, multiple inputs from different people would help capture more complete patterns of causally related functions in natural-language text. Another approach could be to formally represent and evaluate the patterns of causally related functions. This ontological approach would also help other researchers reuse our extraction techniques in other natural-language processing algorithms.

5. CHALLENGES OF THE EXTRACTION APPROACH

This section describes possible improvements to the extraction algorithm to more completely capture causally related functions in biological text.

5.1. Causally related functions from multiple sentences

The main limitation of the current algorithm is that it can only identify causally related functions from a single sentence, as the parser only identifies grammatical relations between words within a single sentence. Identifying causally related functions across multiple sentences would require anaphora resolution, which is a significant challenge being researched in computational linguistics.

5.2. Causally related functions involving light verbs

Section 3.5.1 described the case when a light verb itself does not provide much semantic context, e.g., "to use," but provides a meaningful strategy when combined with an object, e.g., "to use energy." The current algorithm removes any matches containing light verbs.

Computational linguistics applies *light verb construction* to address this problem. The technique reduces a light verb and its following object into a "heavy" verb that has more semantic context on its own, e.g., reduces "to use heat" into "to heat." This construction process looks at whether the object itself can be expressed in verb form. However, the original nuance may be lost in some cases, as in the example the meaning of "to use heat" slightly differs from the meaning of "to heat." Also, for the original example "to use energy," the object cannot be used as a verb, unless a derived verb "energize" is used in which case the original meaning is lost.

5.3. Causally related functions in conjunction

The current extraction algorithm cannot disambiguate causally related functions that involve the conjunction "and." The following sentence from Purves et al. (2001) presents the case when the verbs "cover" and "protect" are related by "and" to imply causality:

> "The exoskeleton extends back from the head to **cover** *and* **protect** other segments."

In many instances, however, two verbs in a conjunction are used to describe the sequence of a particular biological phenomenon, e.g., from Purves et al. (2001):

> "Mineral ions **enter** *and* **move** through plants in various ways."

We are currently looking at techniques to disambiguate these two cases. A possible solution could be to determine whether two verbs in a sentence share the same object. In the first example, both "cover" and "protect" describe the object "other segments." More examination of sentences containing a pair of verbs in conjunction is required.

6. POTENTIAL BENEFITS OF THE EXTRACTION APPROACH

The extraction technique used in this research has important benefits that could enable it to be used for supporting other design research. In addition, causally related functions extracted with the technique could be useful in supporting creative concept generation.

6.1. The extraction technique as a computational tool

6.1.1. Scalability

The extraction algorithm is inexpensive to run. Besides a few manual pre-processing steps, all of the tagging, parsing, and causal-relation extraction processes are performed automatically. Our previous work, e.g., identification of biologically meaningful keywords by Chiu and Shu (2007) and Cheong et al. (2011), required manual processing and therefore had limited scalability. Because the current algorithm is highly automated, it can be used to retrieve information from a large amount of text.

6.1.2. Domain-independence

The extraction approach demonstrated that implicit causal relations, such as causally related functions, can be identified with specific syntactic relations between verbs. Because these syntactic relations are consistent in any English text, our extraction algorithm could be used to extract causally related functions in other domains than biology.

6.1.3. Flexibility in algorithm rules

Because each syntactic pattern defined in Table 2 is mutually exclusive, the algorithm could easily include or exclude a specific syntactic pattern to adjust precision or recall. If an additional syntactic pattern of causal relations is identified, its individual precision/recall scores can be used to determine whether the new pattern should be included as part of the syntactic rules in the algorithm.

6.2. Application of the extraction technique for other design studies

The extraction approach could be used for other design studies that require information extraction and analysis from natural-language text.

6.2.1. Patent data mining

Most relevant to this research is information extraction from patent databases or design documents. Li and Tate (2010) applied natural-language processing techniques to extract functional requirements and design parameters from patents. Fu et al. (2011) used a structure-discovering algorithm on the syntactic data of patents to find structural patterns in the patents. Verhaegen et al. (2011), also applying natural-language processing techniques, derived product aspects from patents that can be used to identify candidate products for design-by-analogy.

Fu et al. and Verhaegen et al. transformed the text descriptions of patents into mathematical representations and evaluated similarity between the representations. This approach is advantageous for handling and computing large amount of data. However, mathematical representations are only approximations to the original text data and can overlook the context involved. On the other hand, our approach could extract important keywords and semantic information in the text descriptions of patents and compare them to find analogous mechanisms in patents.

6.2.2. Knowledge acquisition for functional modeling

A number of structure-behavior-function or function-behavior-structure models (Qian and Gero 1996, Bhatta et al. 1994, Chakrabarti and Bligh 2001) have been developed to model engineering products and compare functional similarities between the products. As discussed in Section 2.1, some of these models are used to represent biological knowledge (Chakrabarti et al. 2005, Goel et al. 2009). All these models try to represent and understand casuality in design knowledge.

An important benefit of the formal and abstract representation of design knowledge is that machines could compare similarities between models of different design knowledge and support analogical design. In the future, more studies could be conducted on automatically converting unstructured, available design knowledge (e.g., in natural-language format) into structured computational models. Because the causal-relation extraction approach discussed in this papar is scalable and domain-independent, the approach may be used to automatically acquire candidate design knowledge for functional modeling.

6.2.3. Design protocol analysis

Computational linguistics is increasingly used to analyze verbal protocols in design. Dong (2004, 2005) used latent semantic analysis to quantify coherent thinking and lexical chain analysis to evaluate concept formation in design teams. Wang and Dong (2008) used statistical patterns of relevant keywords to compute appraisals in design text. In addition, Chiu and Shu (2008) used part-of-speech tags to analyze the degree of functional aspects considered in design concepts.

As another approach to analyze design protocolos, the authors are interested in quantifying analogical reasoning of designers. Cheong et al. (2012) report initial work towards this goal, based on hand-coding different types of similarity comparisons in design protocols. Perhaps the causal-relation extraction technique could be used to help identify higher-level similarity comparisons in design protocols, which may indicate that designers are effectively using analogical reasoning in their concept generation. Because analogical reasoning is thought to be central to creative concept generation, quantifying analogical reasoning could be very useful in evaluating creativity in a design process.

6.3. Causally-related functions in concept generation

The following potential benefits of categorization by enabling functions were generalized based on feedback from search-tool users. Thrty-four students in a fourth-year mechanical design course used the search tool to solve two design problems as a take-home exercise. We also compare these benefits against the search tool developed by the Biomimicry Institute, which is available at the AskNature website (http://www.asknature.org).

6.3.1. Support functional associations in design-by-analogy

Fully understanding descriptions of biological phenomena usually required students to look up additional resources, e.g., Wikipedia, biological dictionaries, etc. Although the extra effort is necessary once a particular biological phenomenon is chosen as candidate analogy, the challenge in understanding terminologies prevents designers from quickly determining the relevance of each search match to design problems. The students suggested that the causally related functions highlighted in search matches helped them associate functional similarities between the biological phenomena and design problems. This observation agrees with findings in cognitive psychology (Gentner 2006), which suggest that similarities found at the functional level guide people to detect the appropriate analogy. AskNature's search tool provides detailed descriptions of a particular biological phenomenon, but designers might find it difficult to quickly make functional associations between the domain-specific descriptions to design problems.

6.3.2. Organize search results by enabling functions

AskNature's search tool returns descriptions of biological phenomena in an unspecified order, requiring designers to examine the relevance of each description one-by-one. Our search tool groups similar biological phenomena by the enabling functions of causal relations. The students reported that this categorization helped them determine which groups of causal relations are more relevant to the design problems.

Our search tool also ranks the enabling functions of causal relations based on their frequency. For example, Figure 1 shows that for the keyword "move," the most frequent enabling function was "beating" performed by cilia and flagella. More ubiquitous strategies in biology may warrant more consideration for design-by-analogy. The grouping of enabling functions also could identify a common strategy that is shared by multiple biological systems.

7. CONCLUSION

The causal-relation extraction algorithm described in this paper uses a set of syntactic patterns to identify causally related functions in biological text. Although implicit causal relations are complex and involve inference based on semantic analysis (Girju 2003), our research demonstrates that a set of syntactic relations can be used to extract causally related functions. The extraction approach is inexpensive, scalable, and domain-independent because the algorithm is based on automatic tagging/parsing and identification of syntactic patterns in text. On the other hand, the performance of the extraction algorithm is limited by the performance of the current state-of-the-art parsing techniques in computational linguistics.

The extraction algorithm further supports the natural-language approach to biomimetic design, which can benefit both designers and researchers. Causal relations categorized by enabling functions help designers identify multiple analogies from various biological phenomena. Researchers compiling the models of biological information could also use the extraction technique to identify candidate biological information. The technique could also be used to extract patterns of ubiquitous strategies in biology from a large amount of natural-language text. These ubiquitous strategies may be applicable to solve problems from multiple domains, further enhancing the potential use of biomimetic design.

ACKNOWLEDGMENTS

The authors acknowledge the financial support of the Natural Science and Engineering Research Council of Canada. We thank Jon Conte for his assistance in developing the extraction algorithm and the search tool.

REFERENCES

Baclawski K, Niu T (2006) Ontologies for Bioinformatics. MIT Press, Cambridge MA.

Bhatta S, Goel A, Prabhakar S (1994) Innovation in analogical design: A model-based approach. Proc. of AI in Design 57-74.

Chakrabarti A, Bligh TP (2001) A scheme for functional reasoning in conceptual design. Design Studies 22(6):493-517.

Chakrabarti A, Sarkar P, Leelavathamma B, Nataraju B (2005) A functional representation for aiding biomimetic and artificial inspiration of new ideas. AIEDAM 19(2):113–132.

Cheong H, Hallihan G, Shu LH (2012) Understanding analogical reasoning in biomimetic design: An inductive approach. To appear in Design Computing and Cognition '12.

Cheong H, Chiu I, Shu LH, Stone R, McAdams D (2011) Biologically meaningful keywords for functional terms of the functional basis. Journal of Mechanical Design 133:021007.

Cheong H, Shu LH (2009) Effective analogical transfer using biological descriptions retrieved with functional and biologically meaningful keywords. Proc. of ASME IDETC2009-86680 (DTM).

Chiu I, Shu LH (2007) Biomimetic design through natural-language analysis to facilitate cross- domain information retrieval. AIEDAM 21(1):45-59.

Chiu I, Shu LH (2008) Effects of dichotomous lexical stimuli in concept generation. Proc. of ASME IDETC2008-49372 (DTM).

de Marneffe M-C, MacCartney B, Manning CD (2006) Generating typed dependency parses from phrase structure parses. In the International Conference on Language Resources and Evaluation.

Dong A (2004) Quantifying coherent thinking in design: A computational linguistic approach. Design Computing and Cognition '04:521-540.

Dong A (2005) Concept formation as knowledge accumulation: A computational linguistics study. AIEDAM 20:35-53.

Fu K, Kotovsky K, Cagan J, Wood K (2011) Discovering structure in design databases through functional and surface based mapping. Proc. of ASME IDETC2011-48322 (DTM).

Garcia D (1997) COATIS, an NLP system to locate expressions of actions connected by causality links. Proceedings of Knowledge Acquisition, Modeling and Management, 10th European Workshop:347-352

Gero JS, Kannengiesser U (2006) A function-behaviour-structure ontology of processes. Design Computing and Cognition '06:407-422.

Genter D (1983) Structure-mapping: A theoretical framework for analogy. Cognitive Science 7:155-170.

Gentner D (2006) Analogical reasoning, psychology of. Encyclopedia of Cognitive Science. John Wiley & Sons, Ltd.

Girju R (2003) Automatic detection of causal relations for question answering. Proc. of the 41st Annual Meeting of the Association for Computational Linguistics.

Goel A, Rugaber S, Vattam S (2009) Structure, behavior, and function of complex systems: The structure, behavior, and function modeling language. AIEDAM 23(1):23-35.

Hacco E, Shu L (2002) Biomimetic conept generation applied to design for remanufacture. Proc. of ASME DETC2002-34177 (DFM).

Joskowiscz L, Ksiezyk T, Grishman R (1989) Depp domain models for discourse analysis. The Annual AI Systems in Government Conference.

Kaplan RM, Berry-Rogghe G (1991) Knowledge-based acquisition of causal relationships in text. Knowledge Acquisition 3(3):317-337.

Ke J, Chiu I, Wallace JS, Shu LH (2010) Supporting biomimetic design by embedding metadata in natural-language corpora. Proc. of ASME IDETC2010-29057 (DTM).

Khoo C, Chan S, Niu Y (2000) Extracting causal knowledge from a medical database using graphical patterns. Proc. of the 38th Annual Meeting of the Association for Computational Linguistics.

Li Z, Tate D (2010) Automatic function interpretation: Using natural language processing on patents to understand design purposes. Proc. of ASME IDETC2010-29097 (DTM).

Mak TW, Shu LH (2008) Using descriptions of biological phenomena for idea generation. Research in Engineering Design 19(1):21–28.

Manning C, Schutze H (1999) Foundations of Statistical Natural Language Processing. MIT Press, Cambridge MA.

Marcus MP, Santorini B, Marcinkiewicz MA (1993) Building a large annotated corpus of English: The Penn Treebank. Association for Computational Linguistics 19(2):313-330.

Nagel J, Nagel R, Stone R, McAdams D (2010) Function based, biologically inspired concept generation. AIEDAM 24(4):521-535.

Purves WK, Sadava D, Orians GH, Heller HC (2011) Life, The Science of Biology 6/e. Sinauer Associates, Sunderland MA.

Qian L, Gero JS (1996) Function-behavior-structure paths and their role in analogy-based design. AIEDAM 10(3):289-312.

Shu LH (2010) A natural-language approach to biomimetic design. AIEDAM 24(4):483-505.

Shu LH, Ueda K, Chiu I, Cheong H (2011) Biologically inspired design. CIRP Annals 765:1-21.

Stone RB, Wood KL (2000) Development of a functional basis for design. Journal of Mechanical Design 122:359-369.

Toutanova K, Manning CD (2000) Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. Proc. of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora:63-70.

Vandevenne D, Verhaegen P-A, Dewulf S, Duflou JR (2011) A scalable approach for the integration of large knowledge repositories in the biologically-inspired design process. The 18th International Conference on Engineering Design.

Verhaegen P-A, D'hondt J, Vandevenne D, Dewulf S, Duflou JR (2011) Identifying candidates for design-by-analogy. Computers in Industry 62:446-459.

Wang X, Dong A (2008) A case study of computing appraisals in design text. Design Computing and Cognition '08:573-592.

Witten IH, Frank E (2000) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers, San Francisco, CA.

