Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Reflective Random Indexing and indirect inference: A scalable method for discovery of implicit connections

Trevor Cohen [a,*], Roger Schvaneveldt [b], Dominic Widdows [c]

[a] *Center for Cognitive Informatics and Decision Making, School of Health Information Sciences, University of Texas, Houston, USA*
[b] *Applied Psychology Unit, Arizona State University, Arizona, USA*
[c] *Google Inc., USA*

ARTICLE INFO

Article history:
Received 24 April 2009
Available online 15 September 2009

Keywords:
Distributional semantics
Literature-based discovery
Implicit associations
Indirect inference

ABSTRACT

The discovery of implicit connections between terms that do not occur together in any scientific document underlies the model of literature-based knowledge discovery first proposed by Swanson. Corpus-derived statistical models of semantic distance such as Latent Semantic Analysis (LSA) have been evaluated previously as methods for the discovery of such implicit connections. However, LSA in particular is dependent on a computationally demanding method of dimension reduction as a means to obtain meaningful indirect inference, limiting its ability to scale to large text corpora. In this paper, we evaluate the ability of Random Indexing (RI), a scalable distributional model of word associations, to draw meaningful implicit relationships between terms in general and biomedical language. Proponents of this method have achieved comparable performance to LSA on several cognitive tasks while using a simpler and less computationally demanding method of dimension reduction than LSA employs. In this paper, we demonstrate that the original implementation of RI is ineffective at inferring meaningful indirect connections, and evaluate Reflective Random Indexing (RRI), an iterative variant of the method that is better able to perform indirect inference. RRI is shown to lead to more clearly related indirect connections and to outperform existing RI implementations in the prediction of future direct co-occurrence in the MEDLINE corpus.

1. Introduction

This paper addresses the issue of indirect inference, finding meaningful connections between terms that are related but do not occur together in any document in a collection. Indirect inference is useful in many applications including information retrieval because documents that do not contain words in a query may be relevant to a user's information need. Thus, retrieval systems that reach beyond query terms can improve performance. Indirect inference is particularly important in the context of developing tools to aid discovery from literature because, by their very nature, discoveries are likely to involve bringing together ideas that have not occurred together previously.

In previous applications, implicit connections between two terms that do not co-occur have been discovered by finding a third bridging term that occurs directly with each of them, according to the discovery paradigm first proposed by Swanson [1]. Several automated methods of knowledge discovery based on this paradigm have been developed and evaluated in the literature [2]. However, given the number of possible combinations of bridging terms and potential discoveries, methods that are able to identify indirect connections without the explicit identification of bridging terms present an attractive alternative. The ability to directly identify implicit connections offers considerable computational advantages on account of the combinatorial explosion that occurs with the number of bridging terms permitted in the chain from cue concept to discovery. Even with only one linking term, the search for a novel discovery requires the following five stages, as described by Yetisgen-Yildiz and Pratt [3]:

1. Terms directly co-occurring with a given starting term are retrieved using a correlation mining approach.
2. A set of these ranked above some predetermined threshold are selected as linking terms.
3. Terms directly co-occurring with each of these linking terms are retrieved using a correlation mining approach, and are selected as target terms.
4. Those terms directly co-occurring with the starting term are excluded (consequently the end result of this process is indirect inference).
5. The remaining terms are ranked using a ranking approach.

This process carries considerable computational and disk I/O expense which limit the possibilities for highly interactive and responsive discovery support tools, unless significant constraints

[*] Corresponding author.
E-mail address: Trevor.Cohen@uth.tmc.edu (T. Cohen).

1532-0464/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2009.09.003

are placed on the discovery search space. As the process does not utilize indirect inference, linking two terms that do not co-occur requires the construction of a path through the discovery space that traverses terms that co-occur directly. This process would be more accurately described as "direct inference" as an explicit pathway from source to target must be established before a discovery can occur. By comparison, corpus-derived statistical models of semantic relatedness such as Latent Semantic Analysis (LSA) [4] are able to identify directly meaningful associations between terms that do not co-occur. For example, in LSA, each term is represented as a vector, and meaningful connections between terms that do not co-occur can be retrieved and ranked using the following process:

1. Retrieve the vector for the starting term.
2. Compare this to the vector for all possible target terms.
3. Exclude those which directly co-occur with the starting term.

The performance of this approach is further enhanced by the condensed nature of the vector space representation—it is possible to maintain the vectors for all possible target terms in RAM, and disk lookup is required for the final step in this process only. LSA is one of several methods provided by the emerging field of distributional semantics that are able to learn meaningful associations between terms from the way in which they are distributed in natural language text. Of these methods, LSA [4] in particular has been shown to make meaningful estimates of the semantic relatedness between terms that do not co-occur directly. This suggests that such models may be useful as means to discover implicit connections in the biomedical literature without the need to explicitly identify a bridging term.

However, the generation of the condensed vector space representations utilized by such models often carries considerable computational cost. LSA, for example, is dependent upon the singular value decomposition (SVD), a computationally demanding method of dimensionality reduction to draw such associations. Consequently, LSA requires computational resources beyond the reach of most researchers to scale to large corpora such as the MEDLINE corpus of abstracts. In this paper we address this issue by evaluating the ability of Random Indexing (RI), which has recently emerged as a scalable alternative to LSA, to derive meaningful indirect inferences from general and biomedical text. We find the original implementation of this method is somewhat limited in its ability to indirectly inference, and propose and evaluate Reflective Random Indexing (RRI), a methodological variant that is customized for this purpose.

The primary motivation of this paper is to demonstrate that the indirect inferencing ability of RI is vastly improved when an iterative approach is utilized. As the scalability advantages of RI are retained, this improvement has significant implications for information retrieval and distributional semantics in general. In addition, we wish to evaluate these models as tools to support literature-based discovery.

The organization of the paper is as follows. Section 2 introduces indirect inference, provides illustrative examples of the ability of LSA to perform indirect inference, and discusses the significance of this ability for the discovery of implicit connections in biomedical text. Section 3 describes RI and its variants, sliding-window (or term–term) based RI, and Reflective Random Indexing, also with illustrative examples of the ability of these models to derive meaningful indirect inferences. In Section 4 we evaluate the ability of variants of RI to simulate Swanson's original discoveries of implicit connections between Raynaud's Disease and dietary fish oil, and migraine and magnesium. In addition, we present a large-scale evaluation of the abilities of these models to derive meaningful indirect inference from a time-delimited segment of the MEDLINE database. A discussion of these results and conclusion follow.

2. Background

2.1. Indirect inference

In the context of distributional models of semantic relatedness, an indirect inference is considered to be a measurable semantic relation between two terms that do not co-occur directly together in the corpus used to generate the model concerned. A simple network model of indirect inference can be generated by considering terms that co-occur in documents to be directly linked. In such a model, indirect inferences correspond to terms that are not directly connected but are connected to the same other terms. With such a model there are various ways to determine the "strength" of the indirect inferences including the number of shared intermediate connections and the strength of the direct connections involved. Paths with more than one intermediate node could also be considered. The important property to be preserved is to discriminate between terms that co-occur and those that do not co-occur but are connected by short paths. Indirect inference is particularly important in the domain of information retrieval, as an information request based on a search term would ideally retrieve related documents that do not contain this term. This issue was a primary motivation for the development of methods such as Latent Semantic Indexing (LSI) [5] that are able to retrieve with accuracy related documents that do not state a particular search term explicitly.

Indirect inference has much in common with the traditional use of "middle terms" in logic, as introduced by Aristotle (Prior Analytics Bk. 1 Ch. 4 and thereafter). For example, in the inference "Socrates is human, humans are mortal, therefore Socrates is mortal", "human" is the middle term, and empirically is the term we would like to find in text to answer the question "Is Socrates mortal?". Middle terms through which inferences are made often appear as bound variables in computational logic, and just as several paths may be chosen through a network, the same formal inference may be made by way of several different functional arrangements of bound variables. In practice, performing any such process by analyzing human language is fraught with difficulty, as Aristotle points out, for example:

> "*It is clear that the middle must not always be assumed to be an individual thing, but sometimes a phrase . . .. That the first term belongs to the middle, and the middle to the last, must not be understood in the sense that they can always be predicated of one another, or that the first term is predicated of the middle in the same way as the middle is predicated of the last.*"
>
> — Prior Analytics, Bk. 1. Ch. 35, 36.

In the context of literature-based knowledge discovery middle terms are generally referred to as "bridging terms" or "linking terms", terminology we will employ for the remainder of the paper. While accepting such complexities as term compounding and ambiguity of relationships, fields such as computational semantics and literature-based knowledge discovery have sought and to some extent found methods for traversing middle terms automatically in ways that can enable the more rapid discovery of potentially interesting connections in scientific information, as will be described in this paper.

2.2. Indirect inference and literature-based knowledge discovery

This capacity to indirectly inference is of particular interest to researchers in the field of literature-based discovery, a field which can trace its inception to the fortuitous discovery of a previously unpublished therapeutic relationship between fish oil and Raynaud's Disease, a circulatory disorder affecting the peripheral vasculature, by Don Swanson [1]. The premise underlying Swanson's