Machine learning-assisted industrial symbiosis
Testing the ability of word vectors to estimate similarity for material substitutions

Chris Davis\(^1\) \quad \mid \quad\) Graham Aid\(^2\)

\(^1\) Invenia Labs, Cambridge, UK
\(^2\) Ragn-Sells Group, Stockholm, Sweden

**Abstract**

A challenge of facilitating industrial symbiosis involves identifying novel uses of waste streams that can satisfy the demands of other industries. For these efforts, a variety of characteristics must often be considered. A mine of relevant knowledge has been gathered in resources such as academic journals and patent databases. However, in looking to harness the potential of such data to support facilitation, compiling information on expansive ranges of material properties and technical requirements from a variety of unstructured sources can pose a significant manual effort. To ameliorate this, we demonstrate and evaluate an automated system that, given a large collection of patents and academic articles related to waste valorization, is able to assist with the process of identifying which waste streams could potentially be used as substitute feedstocks. Instead of aiming to measure (potentially thousands of) material properties directly, we use word correlations as a proxy to reflect “common knowledge.” Novel in furthering this approach is the application of word vectors, which have emerged as a promising natural language processing tool. The process employs a machine learning approach where words are represented as high-dimensional vectors which encode latent features related to words that often appear around it. When this approach is assessed by comparing its suggestions to documented cases, the use of vectors shows potential to incorporate latent information in data-based explorations. Further research into how this approach compares, and could be integrated with, established symbiosis development practices will be key to understanding its full potential and drawbacks.

**KEYWORDS**

industrial ecology, industrial symbiosis, machine learning, resource recovery, tacit knowledge, word vectors

**1 INTRODUCTION**

A challenge of industrial symbiosis (IS) and waste valorization involves identifying novel uses of waste streams which can satisfy the demand for feedstocks needed by other industries. As the experience and the documentation from the IS community increases through the creation of more successful exchanges (Neves et al., 2020), it is desirable to be able to leverage the growing body of knowledge in IS literature to improve both the diversity and the relevance of potential exchange opportunities in the future.
TABLE 1 Different ways of encoding the same underlying information about a waste material exchange

| Information encoding for:       | Producing industry | Waste stream | Receiving industry |
|--------------------------------|--------------------|--------------|--------------------|
| Matching on names              | Pharmaceuticals    | Yeast extract| Salmon farming     |
| Matching on classifications    | NACE 21.2          | EWC 02 03    | NACE 3.22          |
|                                 | (Manufacture of pharmaceutical preparations) | (wastes from fruit, vegetables, cereals, edible oils, cocoa, coffee, tea, and tobacco preparation and processing; conserve production; yeast and yeast extract production, molasses preparation and fermentation) | (Freshwater aquaculture) |
| Matching on explicit properties| Yeast extract composition: | Salmon feed composition: | |
|                                 | ∙ Protein: 50%     | ∙ Lipids: 20% | |
|                                 | ∙ Carbohydrates: 5%| ∙ Protein: >45%| |
|                                 | ∙ Lipids: <1%      |              | |

There are a variety of ways in which this information can be re-used, as matchmaking efforts that link waste streams to re-use opportunities take different forms such as in-person facilitations or automated systems (Álvarez & Ruiz-Puente, 2016; Cutaia et al., 2011; Grant et al., 2010; Trokanas et al., 2014; van Capelleveen et al., 2018). While years of experience may improve the process of in-person facilitations, there is still a question of how this could be augmented by more automated methods that could scan large amounts of literature and provide useful suggestions for IS practitioners.

In this article, we demonstrate a method using machine learning and big data that looks to facilitate the process of identifying relevant potential exchanges. Particularly of focus in this work is exploring how such systems might function outside the constraints of taxonomies (e.g., not relying on tag or name matches in identifying exchange potentials). While this approach is not meant to replace existing efforts, we believe that such taxonomy-unbound approaches can certainly augment existing systems and lead to the development of hybrid systems which leverage the strengths of both humans and computers.

2 | BACKGROUND TO IS MATCHMAKING AND THE ENCODING BOTTLENECK

2.1 | Encoding information for automated IS matchmaking

In simplistic terms, most matchmaking systems can be thought of as comprising three steps:

- Observe information about known waste exchanges between industries
- Encode information about these observations so they can be recorded in some data storage system (a notebook, spreadsheet, database, etc.)
- Retrieve and process this information in a way that allows for useful predictions of waste re-use opportunities between organizations.

Systems that facilitate matchmaking differ significantly in their approaches based on how that information is encoded and how that encoded information is then processed to provide recommendations (van Capelleveen et al., 2018). Table 1 shows several common options for encoding the same information for matchmaking. In this example, a pharmaceutical company produces yeast extract as a waste flow, which is then re-used in feed for a salmon farm. As described below and in Table 2, there are several advantages and disadvantages for each approach.

2.2 | Matching on names

First, one may create a simple spreadsheet table that lists which waste streams can be re-used by which industrial processes. This approach is straightforward, but its drawback is that it does not facilitate “novel sourcing of required inputs” (Lombardi & Laybourn, 2012), beyond those exchanges already explicitly mentioned in the table. While it is useful to suggest exchanges which are already known to be feasible, researchers are also interested in matchmaking systems that can help to identify new types of exchanges which are not explicitly documented or previously considered (Chertow & Park, 2016).
### TABLE 2  Advantages and disadvantages of different information encoding strategies for IS matchmaking

| Technique                  | Advantages                                                                 | Disadvantages                                                                 |
|----------------------------|-----------------------------------------------------------------------------|-------------------------------------------------------------------------------|
| Matching on names          | • Very simple                                                               | • Predictions based on exact matches, does not help with finding novel opportunities  |
|                            |                                                                            | • Must deal with synonyms or related terms                                    |
|                            |                                                                            | • Does not generalize insights                                                 |
| Matching on classifications| • Provides standardized names                                             | • Categories can be too general                                                 |
|                            | • Allows for links to statistical data (IO tables, facility databases)      |                                                                                |
| Matching on explicit properties | • Generalizes insights that can be extrapolated to other waste streams and processes | • Data collection can be prohibitively expensive                              |

#### 2.3  Matching on classifications

An alternative approach is to perform matching using existing classification systems (Aid et al., 2015; Costa & Ferrão, 2010a; Massard & Erkman, 2009), instead of matching based on the names of wastes and industrial processes. As shown in Table 1, one could encode industries using the NACE industrial classification codes and encode information about waste streams using the European Waste Catalogue (EWC).

An example of this approach has been done by Costa and Ferrão (2010a) who surveyed examples of industrial symbiosis, summarizing the number of exchanges seen between companies of particular NACE codes. This work can be used as a basis for a matchmaking system, by first locating pairs of NACE codes for which an exchange is known to exist, and then examining new industrial ecosystems and finding pairs of companies having the same combinations of NACE codes.

An advantage of this approach is that it allows us to link to existing statistical data that can help with locating facilities (via pollution databases such as the European Pollutant Release and Transfer Register or the US Toxics Release Inventory) or understanding the size of flows for an industrial sector within a country (PRODCOM and national input–output tables).

As seen in Table 1, a disadvantage is that we are at the mercy of the classification system, which has not necessarily been designed to meet the level of detail that we are interested in. For example, the EWC code for yeast extract also includes a wide array of organic wastes with differing properties.

#### 2.4  Matching on explicit properties

A more flexible approach would be to recognize that we need to match properties of waste streams with the properties of feedstocks required by industrial processes. Using this approach, one would create a database documenting properties of waste streams such as energy density and chemical composition while also doing the same for the properties required for feedstocks. These properties may be a single number (i.e., boiling point at standard temperature and pressure), a range of observed values (higher heating values for waste biomass streams), or limits on acceptable values for feedstocks (maximum sulfur content). When new waste streams are added, their measured properties can be added as well, and we can then automatically identify which processes could potentially accept this new feedstock, based on the waste stream's defined properties. While this approach is more robust than the first, there can still be significant overhead when adding new waste streams and feedstock descriptions due to the need to measure properties or find their values in literature. Furthermore, one may find that the list of required properties for determining the feasibility of a match may become excessive and expensive to compile, leading to questions about which properties are most beneficial to measure for which items. One may also create extensive rule-based expert systems, although Grant et al. (2010) have noted that the approach “is less than elegant and can quickly balloon into a seemingly inexhaustible and arbitrary alchemy of rule-based methods.”

In recent years, tools have been further developed to improve upon and incorporate several of the above strategies for matchmaking systems. For example, new approaches to harnessing the semantic web to increase the usability and interoperability of broader data sources (e.g., data on properties, sources, location, etc.) have been advanced (Ghali & Frayret, 2019). Furthermore, efforts have been made in developing hybrid approaches to elicit more exchange potential through less explicit information. One example being the case-based reasoning approach (proposing solutions based on those that worked for situations with similar aspects) taken by Gatzioura et al. (2019) in their development of the Sharebox platform.

Still, a difficulty faced by any automated IS system is that the identification of relevant exchanges can involve large amounts of both explicit and tacit knowledge and that the encoding of this information into an automatic system can pose significant challenges, as described by Grant et al. (2010):
The taxonomical classifications of resources are at present a great challenge to ICT search tools. [...] For instance, cardboard and paperboard may be substitutable or identical inputs for a by-product process, but their equivalency is based on a more tacit knowledge, which is not easily coded into a computer system. Similarly, resources like “waste water” require an enormous list of attributes for a computer to establish an acceptable match. [...] Input-output matching is at this stage exceedingly difficult to codify and thus relies on communication methods more suited to tacit knowledge.

2.5 | The encoding problem

Each of these approaches uses different strategies to encode (i.e., record) observations in a way that captures information that is later useful for prediction, and as noted, each of these has different tradeoffs. Ideally, we want to be able to encode our observations in a format that captures generalized insights that can help identify multiple diverse re-use opportunities. In other words, knowing that yeast can be used as salmon feed is not as insightful as being able to understand the context behind that observed exchange.

For example, what makes the yeast interesting for this exchange is that it has a high protein content, meaning that other waste streams with a high protein content could be interesting for exchanges as well. On the receiving end, the generalized insight is that waste streams with high protein content are useful as animal feed, and not just for salmon. This means that whenever an IS practitioner spots a waste stream with a high protein content, they should explore possible connections with both aquaculture and land-based agriculture. Understanding the underlying reasons for an exchange can give practitioners a much larger permutation space of potential re-use opportunities. At a higher level, for an initial matchmaking scan, we are not so concerned about searching for a specific underlying reason but, for the sake of discovering novel exchanges, are open to any underlying reason that could enable a potential exchange, even if these reasons are not immediately obvious to us.

So how can we identify and encode these underlying properties and insights? Many of the problems faced by traditional approaches to IS matchmaking can be summarized by Yann LeCunn, the director of Facebook’s AI research group and a pioneer in the field of neural networks, when he states that “The problem is that language is a very low-bandwidth channel. Much information that goes through language is because humans have a lot of background knowledge to interpret this information” (Simonite, 2017).

This leaves us in a situation where, while there is a large and increasing amount of literature available on waste materials and industrial processes through sources such as academic articles and patent databases, we face an encoding bottleneck constrained by people’s ability to encode this information in a way that can be sensibly used in an automated system. The reason we claim there is a bottleneck is because material databases, taxonomies, classifications, and ontologies are generally created by filtering and summarizing the original source literature. This means that there is a large amount of information left behind that could be potentially relevant but was discarded in the process of summarization due to the use cases of those who compiled these. The essence of the encoding bottleneck problem is that if we summarize an encyclopedia down to a small book, no amount of machine learning or other clever approaches can retrieve information that has been discarded when summarizing the original source. This fundamentally limits the potential of any downstream applications which use this data. While we desire to capture as much potentially useful information as possible, we face a tradeoff based on the cost of encoding more information versus the potential improvements to downstream applications.

A seemingly rational way to approach the automation of encoding information would be to have a computer program that could read all of this literature and then automatically extract information on waste streams, feedstocks, and all their various relevant properties. In other words, have a computer do what people have been doing by hand, but just much quicker. The challenge of this approach is that it is still very difficult using computerized natural language processing (NLP) methods to reliably extract this information from complex text such as that found in academic articles and patent documents. We propose in this paper a different approach that, while not able to retrieve measured values of properties, is capable of processing large amounts of text with minimal human intervention, while being able to provide useful recommendations for potential IS exchanges.

2.6 | Conceptualizing wastes and feedstocks as vectors

To understand the approach we propose, it is first useful to conceptualize waste streams and feedstocks as vectors, where the dimensions of these vectors represent a multitude of aspects, such as suitability for applications, general properties, or chemical composition. The top section of Figure 1 gives an illustration of how characteristics of a waste stream (e.g., “high compressive strength,” “low moisture content,” and “combustible”) could be represented as dimensions of a vector, where darker colors indicate higher scores or values for those properties. For the sake of discussion, the exact values are not important, but rather if they are generally high or low.

The middle section of Figure 1 shows an equivalent illustration for feedstocks. When considering if a feedstock is suitable for a process, we can examine the vector for a feedstock in terms of required properties and incidental properties. We define required properties as the properties that must be satisfied for the feedstock to be a suitable input for the process. Incidental properties, however, are properties of the feedstock that are irrelevant for the process, and the presence or absence of these has no influence on determining the feedstock’s suitability. For example, meat and bone meal can be used as a fuel in a power plant because it is combustible, and its high protein content is irrelevant for this consideration.
Performing matchmaking activities to re-use a waste stream from one industry as a feedstock for another can be thought of as an exercise in comparing vectors. Specifically, we are comparing the vector representing a waste stream with the vectors for feedstocks, with our comparisons focusing on the dimensions representing the required properties. As shown in the bottom section of Figure 1, we may find that a waste stream is suitable for multiple processes, but for different fundamental properties.

What is interesting about representing waste streams and feedstocks as vectors is that it allows us to numerically compare them. Measuring the angular distance between vectors (commonly referred to as the cosine distance) gives us a single numerical value indicating how similar those two vectors are. Considering the case shown at the bottom of Figure 1, the cosine distance between the waste stream vector and the vectors of both feedstocks may be the same, even though they are similar with regard to different dimensions:

\[
\text{cosine similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|} \tag{1}
\]

\[
\text{cosine distance} = 1 - \text{cosine similarity} = 1 - \cos(\theta) \tag{2}
\]

There are several key points from this example. First, creating recommendations for potential IS exchanges is a process that involves comparing the similarity of items across multiple dimensions. Second, a simple numerical measure of similarity between these items (i.e., waste streams and feedstocks) may be a useful indicator that can contribute to the identification of potential IS exchanges, as will be further explored in this paper.

This strategy has been described by Trokanas et al. (2014), who compare materials based on vectors of their properties, and further augment this comparison by finding the shortest distance between these two materials within a graph defined by an ontology of materials. The limitation of this approach is that there is an encoding bottleneck—the creation of this database moves at the speed at which a human can encode information into this system. One has to determine which material properties may be relevant for matchmaking, and then be able to locate suitable measurements for them. This process can be very tedious and expensive, especially for large numbers of diverse materials. Such bottlenecks and drawbacks lead us to the questions we would like to address in the current research.
3 | RESEARCH QUESTIONS

Taking someone who is interested in facilitating IS by comparing a multitude of potentially relevant properties (the diversity of which may be too numerous to explicitly and comprehensively list) as our point of departure, in this paper, we investigate the following research questions:

1. How could one efficiently and usefully encode diverse properties of wastes and feedstocks through the use of automatically generated vectors and apply this to IS facilitation?
2. What approach could be taken to measure the accuracy of novel IS matchmaking systems?

4 | METHODS

In this section, we describe machine learning approaches for overcoming the encoding bottleneck before demonstrating their use in the novel matchmaking approach. The previous quote from LeCunn (Simonite, 2017), about how we understand language through background knowledge of the words involved means that we need a way to be able to associate a particular term, like "cardboard" or "olive mill sludge," with background information that can help describe it and give context. Some of the first insights on how to approach this can be traced to Harris (1954) and also Firth (1957) who stated that "a word is characterized by the company it keeps," meaning that one can learn about a word by examining the words that commonly occur around it in text.

4.1 | Analyzing co-occurrences of terms

There are several ways in which this can be done. A straightforward approach would be to take a large body of text, split it into sentences, and then compile statistics on the number of times that combinations of terms appear together in the same sentence. For example, if one were to analyze a collection of articles returned by an academic literature search for "bio-energy," we would expect the terms "biogas" and "manure" to occur together quite frequently in sentences. The high co-occurrence count value for this combination of terms indicates the presence of what can be termed a "latent property." The property is latent since we know it exists, since something is causing "biogas" and "manure" to frequently occur together in text, but the nature of the actual property (that manure is a feedstock for biogas production) is hidden from us, since we are simply counting co-occurrences of terms. Similarly, "soybeans" and "sunflowers" would likely co-occur frequently with "biodiesel" and "edible oil." If we think about what these relations represent, the fact that a plant can be used to produce an edible oil means that due to the chemical compounds in that oil, it likely also be used as a feedstock for biodiesel production. The presence of these chemical compounds is encoded as a latent property by high co-occurrence counts with "biodiesel" and "edible oil." In general, when we see terms frequently co-occurring, it often encodes information about deeper underlying properties, and that, while these properties are not necessarily explicit, they still give us useful information. The presence of relations between terms often imply other relations as well.

It should also be noted that although we previously discussed required and incidental properties in Section 2.6, examining co-occurrences does not explicitly consider these, although it can be assumed that if a term for a feedstock and an application are frequently seen together, then we are likely seeing some required properties satisfied.

This approach of encoding latent rather than explicit properties helps us to overcome the encoding bottleneck identified in Section 2.1, and the experimental methodology detailed later in this article demonstrates that we are able to predict latent relations between terms which were not explicitly seen in the source text, but were inferred by a collection of co-occurrence of other terms. This ability to encode latent properties enables the automated processing of very large amounts of text relevant for IS practitioners, although with the trade-off that we must then use other techniques to later determine the explicit properties behind these latent properties. The rest of this article builds upon this insight, while showing more sophisticated approaches and evaluating their performance in recommending potential types of IS exchanges that are known to exist from actual case studies.

The limitation of simply counting the co-occurrences of terms is that there will be a natural bias toward the terms which appear the most in the text. We are not interested in the popularity of terms, but rather a measure of how closely associated two terms are. A multitude of statistical measures may be used for this problem (Pecina, 2005), with one such metric being the pointwise mutual information (PMI) (Bouma, 2009), shown in Equation (3). The variables $x$ and $y$ refer to any two terms, such as "fly ash" and "cement production." The probabilities $p(x)$ and $p(y)$ are calculated by counting the number of documents each term appears in, divided by the total number of documents being analyzed. Similarly, $p(xy)$ measures how frequently both terms $x$ and $y$ are found mentioned together in a document. The ratio defined by PMI is then an indicator of how often terms appear together ($p(xy)$), in relation to how often they would be expected to occur together by chance ($p(x)p(y)$). If two terms occur much more than would be expected by chance, then this indicates that there is some sort of relation between them. The PMI can be normalized to values between...
Representing words as vectors

According to Section 2.6 of representing terms for wastes and feedstocks as vectors, we can use the NPMI values for combinations of terms for such a purpose. For example, for the term "biogas," a vector can be created where each dimension contains its NPMI scores when combined with "manure," "brewery waste," and "anaerobic digestion." The limitation of this approach is that it still faces a human bottleneck in terms of determining which dimensions (e.g., properties or terms) are relevant to include and to measure. Furthermore, the idea of using surprising or novel combinations for IS hints that we are interested in latent properties (i.e., properties not immediately obvious to the practitioner, but whose presence could indicate a feasible combination).

4.2 Representing words as vectors

Over 20 years ago, Lund and Burgess (1996) pioneered an approach using co-occurrence statistics to create vectors that aimed to capture information about word meaning. The length of the vectors was based on the total number of unique words in the text examined, meaning that every location in the vector corresponded to a specific term. Reading across the vector for a single word, one would encounter co-occurrence statistics for that word with every other word it potentially co-occurred within the text. Lund noted that the vectors for road and street were similar, as were vectors for coffee and tea. Furthermore, he found that when locating the word closest to another word, one would find that semantic relations were encoded ("jugs" is closest to "cans," "cardboard" is closest to "plastic") along with associative relations ("lipstick" and "lace," "monopoly" and "threat"). The vectors also seemed to encode categorical information. For example, in one experiment, he took a set of vectors corresponding to groups of words describing animals, body parts, and geographical locations. He then reprojected the words from their high-dimensional vector representation to two-dimensional vectors, which were then visualized in the form of a simple x-y scatter plot. The resulting plot could then be divided into sections that partitioned the words into their original defined groups of animals, body parts, and geographical locations.

Lund hints at a way of dealing with the encoding bottleneck previously discussed in Section 2.1 and shows how it is possible to have a computer process unstructured text (i.e., academic articles, news items, books, etc), make observations about which terms are commonly seen with other terms, and then generate a vector representation for terms that encode information that captures properties of the term itself. A drawback of Lund's approach is that it was computationally expensive and not very suitable for processing large amounts of text (i.e., gigabytes of raw text), which could be useful for improving the vector representations of terms.

A major advance was by Mikolov et al. (2013), whose creation of the Word2Vec algorithm was significantly faster than other approaches. Mikolov demonstrated that while previous approaches may take weeks or months to generate word vectors from a certain collection of text, he could achieve similar results within hours.

The first step in this approach is to create the list of terms (called the "vocabulary") for which word vectors will be generated. This is done by first scanning through the input text and keeping only those terms which occur more than a user-defined minimum amount. If one sets this user-defined limit very low (i.e., keep words that occur two times or more in the entire text), then this will result in a very large vocabulary. Since very specific terms such as "blast_furnace_sludge" are likely to occur far less than generic terms like "sludge," this may seem like a good outcome. However, the fewer times that a term appears in the text, the less information that is available for the algorithm to try to encode the term's context based on the words often found nearby it. In other words, the word vectors for terms which appear frequently will likely better encode context than those for words which appear infrequently.

The algorithm then initializes randomized vectors for all the words in the vocabulary. Figure 2 gives a simplified explanation of this by showing vectors for terms initialized in a two-dimensional space. Next, it examines the input text line-by-line, and using a neural network, it adjusts the vectors for terms appearing within a certain user-specified window (number of terms) of each other, in a way that reduces the vector distance between them. As shown in the second panel of Figure 2, every time that two terms are seen in text together, their corresponding vectors are nudged closer toward each other. Over time (the third panel of Figure 2), the vectors tend to coalesce into groups of similar terms.
The size of the vectors is user-specified, and it is common for a value between 300 and 1000 used on word vectors that have been created using large sources of text such as Wikipedia and a dataset based on Google News. The dimensions of these vectors can be seen as encoding latent properties rather than explicit properties. In other words, it is not possible to look at a specific dimension in a word vector and state that it encodes a value related to “heating value” or “carbon content.” Figure 2 illustrates this with the example of two-dimensional vectors. While the clustering of vectors seems reasonable in its ability to group together similar terms, it is difficult to ascribe an exact meaning to the horizontal and vertical axes. The axes are not an attempt to measure an explicit property, but rather, the two dimensions give us a space in which we can rearrange the initially randomized vectors so that they are close to the vectors for other similar terms.

Furthermore, from Figure 2, we can see that it is difficult to encode complex relationships within only two dimensions. While wheat and maize are similar since they are both crops, they also are related to fertilizer. Maize also relates to biogas as it is sometimes used as a feedstock in combination with manure. Fly ash is an ingredient in cement and fertilizers, although cement and fertilizers are not very related themselves. The use of far more than two dimensions in vectors (i.e., 300 or more dimensions) makes it easier to capture these types of relations such as where Y is similar to X and Z, but X and Z are not similar.

In Section 4.4, we show that this abstract representation encodes information which is useful for predicting potential IS exchanges.

4.2.1 Applications of word vectors

Word vectors are currently finding uses in many different applications. Asgari and Mofrad (2015) used an approach based on Word2Vec to create vectors for biological sequences related to proteins and genes. Minarro-Gimenez et al. (2014) used word vectors to help identify the properties of pharmaceuticals. Ghosh et al. (2016) used them as a means to help create a disease taxonomy based on analyzing unstructured text from news sources. By comparing the distances between word vectors, Ghosh was able to identify linkages between diseases and their symptoms, transmission methods, and transmission agents. Others have used word vectors for sentiment analysis of text (Tang et al., 2014) and for characterizing foods and meals (Tansey et al., 2016). Kuczynski et al. (2016) demonstrated how Word2Vec could be used to assist in matching the names of life cycle inventory (LCI) flows from different databases, using the process specified by Kusner et al. (2015). In the example shown, this technique was able to link an LCI flow name mentioning “softwood” to other flows from another LCI database mentioning “cedar,” “spruce,” and “pine,” which are indeed

---

1 https://code.google.com/archive/p/word2vec/, https://nlp.stanford.edu/projects/glove/, https://github.com/facebookresearch/fastText/blob/master/pretrained-vectors.md
4.3 Data preparation

More details about the process of creating the word vectors are provided in Section S.3. At a high level, we first compiled a large collection of source literature on waste re-use from academic literature and also a patent database, resulting in 2.5 GB of raw text. Further pre-processing of this text was done to indicate terms consisting of multiple words (e.g., “wood waste” became “wood_waste”). Finally, this processed text was then used as input to fastText for the creation of the word vectors.

4.4 Comparison against known EIP examples

In this section, we demonstrate how word vectors can be used for IS matchmaking, and evaluate how well this approach can predict known exchanges in existing eco-industrial parks (EIPs). This system recommends potential IS exchanges by using the similarity measurements between terms representing waste streams and also receiving industries. As mentioned previously, there are several techniques which can be used to generate these similarity measurements. For this example, we will use two different approaches (NPMI and cosine similarity of word vectors), both of which use the same pre-processed text described in Section S.3.2. Our main point in this exercise is to quantitatively show the value of using latent relations instead of only explicit relations. There are several types of tasks which can use word vectors for input, such as the similarity matrix in Figure 3.
The reason for comparing NPMI to word vectors is that they differ in the amount of information that is incorporated in their similarity calculations. NPMI calculations rely on scanning for exact matches of two terms in a large body of text. As such, the rest of the terms in the text have no impact at all on the NPMI calculation. For word vectors, this is not the case, due to how the word vectors are generated. This is because the word vector for a term is influenced by the vectors for other terms appearing nearby in the text, which are in turn influenced by other nearby terms. Thus, when we compare the vectors for two words, we are also, to some extent, comparing information related to the "context" or other words that often occur with them as well. While this may be theoretically interesting, if word vectors are able to encode "useful" latent properties in practice, then we should see that they perform better than calculations involving simpler techniques such as NPMI. Furthermore, the comparison with NPMI is interesting as Levy and Goldberg (2014) have argued that word vectors are a means of approximating PMI measures.

The system we demonstrate consists of a binary classifier that gives a simple yes/no decision on recommending an exchange, based on whether the similarity between a waste term and a receiving industry is above a discrimination threshold or not. One of these classifiers uses the NPMI data, while the other uses the cosine distance between word vectors. It should be noted that while both data sets provide similar types of measures, it is not necessarily meaningful to compare both classifiers at the same discrimination threshold. As the overall objective is to produce an accurate classifier, the best performing NPMI-based and word vector-based classifiers will likely use different discrimination thresholds.

To evaluate these, we use information from EIP case studies to measure how well we can predict the actual existing exchanges. The information on actual exchanges comes from the work of Li (2014) who has collected symbiosis exchange information from several literature sources on the following EIPs: Brownsville (Martin et al., 1996), Chamusca (Costa & Ferrão, 2010b), the Finnish forest industry (Pakarinen et al., 2010), Guitang (Zhu et al., 2007), Kalundborg (Kalundborg Symbiose, 2021), Kwinana (Sinclair Knight Merz, 2007), Lubei (Yang et al., 2004), Nanhai (Yang & Feng, 2008), Norrkoping (Baas, 2010), Shanxi (Ding & Hua, 2012), Styria (Schwarz & Steininger, 1997), TEDA (Shi et al., 2010), Tiexi (Tiejun, 2010), and Ulsan (Park et al., 2008; Park & Behera, 2014).

4.5 Preparation of EIP data

The EIP case study data consist of tabular data with four columns describing the exchanges found in the literature:

- Supplier: The industry producing the product or waste.
- Accepter: The industry receiving the product or waste.
- Transaction: The product or waste transferred from the supplier to the accepter.
- Attribute: Description of the nature of the transaction: "Product," "By-product," or "Utility Sharing."

For the matchmaking test, we only use the columns accepter and transaction, for rows where the attribute is "By-product," meaning that we only examine waste exchanges. For each of the different terms describing acceptors and transactions, we located their corresponding word vectors. As described below and in more detail in Davis and Aid (2021), we were able to locate a mix of exact matches, close matches found by hand, and in some cases, no matching word vector could be located.

- Exact match—EIP data: "Cardboard" matched to the word vector for "cardboard."
- Close match—EIP data: "(NH4)2PO4 Production" matched to the word vector for "ammonium_phosphate."
- No match—no corresponding word vector found, often relates to specific company names mentioned or to vague terms being used.

While this dataset is useful to demonstrate a proof-of-concept of the ideas we present, there are several issues with it that should be noted. The case study data contains terms with a wide range of specificity (i.e., "waste_polystyrene" versus simply "ash" or "effluent"). Furthermore, in some cases, we matched on actual company names (Air Liquide, for the data on Kwinana 2000 & 2005). In the case of Air Liquide (a manufacturer of industrial gases), most of the recommendations are indeed on industrial gases, indicating that the word vectors have (in some cases) encoded information about company operations.

5 RESULTS

The results of the analysis are shown in three steps. First, we show the NPMI and word vector cosine similarity matrices, which contain recommendation scores for waste exchanges in EIPs. Second, we examine the highest scoring recommendations which do not match known exchanges (i.e., potentially "novel" recommendations). Finally, we perform a statistical analysis in order to better understand the quality of the recommendations. The results for the rest of the EIPs are included in Supporting Information Sections S.4 and S.5, although for the sake of brevity, we only examine the results for the TEDA EIP here.

Available as an R package at https://github.com/complexly/EIPs
**Table 3** Top exchange recommendations not found among verified exchanges for TEDA. The comments indicate the actual nature of the combination of suggested waste stream and receiving industry. Note that “Receiving Industry” is sometimes in reality the producing industry in cases where “Output of process” is noted. Information regarding underlying data for this table is available in Section S.2 of the Supporting Information.

| Waste stream          | Receiving industry | Comments              | Cosine similarity |
|-----------------------|--------------------|-----------------------|-------------------|
| Steel scrap           | Cast iron          | Recycling             | 0.597             |
| Lead slag             | Lead recycling     | Recycling             | 0.522             |
| Effluent sludge       | Wastewater treatment plant | Treatment        | 0.495             |
| Aluminum scrap        | Cast iron          | No clear connection   | 0.416             |
| Fly ash               | Cement mill        | Re-use                | 0.413             |
| Electronic waste      | Lead recycling     | Recycling             | 0.403             |
| Electroplating sludge | Aluminum electrolysis| Output of process    | 0.400             |
| Spent lead acid batteries | Aluminum electrolysis | No clear connection | 0.396             |
| Galvanized scrap      | Lead recycling     | No clear connection   | 0.385             |
| Lead slag             | Aluminum electrolysis| No clear connection  | 0.380             |
| Electroplating sludge | Wastewater treatment plant | Treatment | 0.379             |
| Electroplating sludge | Cement mill        | Re-use                | 0.370             |
| Aluminum scrap        | Lead recycling     | No clear connection   | 0.368             |
| Steel scrap           | Aluminum electrolysis| No clear connection | 0.361             |
| Steel scrap           | Lead recycling     | No clear connection   | 0.360             |
| Food residues         | Aerobic composting | Re-use                | 0.359             |
| Electroplating sludge | Lead recycling     | Recycling             | 0.351             |
| Flue gas              | Cement mill        | Output of process     | 0.346             |
| Galvanized scrap      | Rubber product     | No clear connection   | 0.343             |
| Sewage                | Aerobic composting | Treatment             | 0.343             |
| Food residues         | Wastewater treatment plant | Treatment | 0.340             |
| Galvanized scrap      | Aluminum electrolysis| No clear connection  | 0.339             |
| Fermentation residues | Wastewater treatment plant | Treatment | 0.315             |
| Fermentation residues | Cogeneration plant | Re-use                | 0.315             |
| Lead scrap            | Aluminum electrolysis| No clear connection  | 0.314             |
| Fermentation residues | Cement mill        | Re-use                | 0.307             |

**NPMI and cosine similarity matrices.** The matrices in Figure 3 show the recommendations for the TEDA EIP, where rows are terms corresponding to waste streams and columns represent terms related to the receiving industries. Darker cells in the matrices indicate a “stronger” recommendation. Cells with dotted red lines indicate exchanges that are actually occurring at TEDA.

As seen on the left side of Figure 3, the approach which uses the NPMI statistics for co-occurring words actually misses all of the verified exchanges, at least in the case of the TEDA EIP. Most of the cells in the matrix are empty, indicating that the very few of the words on the rows and columns were seen mentioned together in the source literature. This demonstrates a key limitation with matching on names, as discussed back in Section 2.1, specifically that of dealing with synonyms and related terms.

The right side of Figure 3 demonstrates the strength of the approach, which calculates the cosine similarity between pairs of word vectors, namely that it allows us to calculate similarity between any pairs of words, even if they do not explicitly appear together in the source literature. Furthermore, most of the verified exchanges receive high scores.

**Examining highest scoring non-verified recommendations.** Table 3 examines the top recommended exchanges for TEDA (with a cosine similarity $\geq 0.3$) from Figure 3 that are not part of the actual verified exchanges. In other words, this shows us potentially valid exchanges that have not yet been realized within the EIP. A limitation of the recommendation approach is that although our intent is to predict exchanges with a specific direction (i.e., a waste stream as an input to a particular industrial process), in reality this approach is only giving an indication of the similarity of the two words being compared. As a result, several exchanges are labeled as “Output of process”, indicating that this exchange is not about waste re-use but rather waste production. Other exchanges are labeled as “Recycling”, “Re-use”, or “Treatment” to describe the nature of the processing done by
the receiving industry. Finally, “No clear connection” is indicated if this does not appear to be a valid exchange. In this case, the pairs of words are all similar in that they relate to the metal industry.

What we see from this example in Table 3 is that by using the cosine similarity between pairs of word vectors describing wastes and receiving industries, we can identify feasible waste exchange opportunities. While this approach does address some problems of the encoding bottleneck described in Section 2.5, care should be taken in interpreting the results as this measure of similarity captures a wide variety of similarities, while for IS purposes, we are interested in a specific type of similarity (i.e., some waste is similar to a feedstock for some industry).

Statistical analysis of recommendations. In order to more systematically evaluate the suggestions proposed by this system, we use several metrics. One distinction to be aware of is that this analysis is based on verified exchanges and not on feasible exchanges. As a result, the classifier results may be penalized despite giving “good advice” because an exchange that may be feasible is not seen within that particular EIP.

Two common metrics used to evaluate classifier results are precision and recall, shown in Equations (5) and (6) where our binary classifier’s yes/no decisions are characterized as true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) based on what is known from the case study exchanges. Precision is a measure of, given the set of recommendations, what percentage of those recommendations are true based on the known exchanges from the case studies; recall measures, given the set of all known verified exchanges from the case studies, what percentage of them actually show up in the recommendations:

\[
\text{Precision} = \frac{TP}{TP + FP} \quad (5)
\]
\[
\text{Recall} = \frac{TP}{TP + FN} \quad (6)
\]

This distinction is important since within an industrial area, the number of possible waste streams and receiving industries results in a potentially large number of permutations, where most of the combinations will likely not be interesting. In other words, the number of negatives will be far larger than the number of positives. This means that if we pick combinations of wastes and industries at random, we are far more likely to generate FP (recommendations that are not verified exchanges) than FN (verified exchanges not recommended), thus resulting in a high recall value and a low precision value according to the definitions of Equations (5) and (6). Furthermore, there tends to be a tradeoff between precision and recall (Buckland & Gey, 1994) where adjusting an algorithm to increase one will often decrease the other.

The left plot of Figure 4 shows the precision measurements for binary classifiers using the data from NPMI calculations and from the cosine distance of word vectors. What this shows is that using NPMI values results in much higher precision. The curve for the NPMI-based classifier stops around 0.6, as that is the value at which the discrimination threshold exceeds the highest NPMI value, thus resulting in no more TP or FP. The high value at the end of the curve is due to recommending a single exchange that also shows up in the case studies. Therefore, in suggesting only a single potential exchange, it is 100% accurate, but also fails to make other predictions that could be useful. As seen, the classifier using word vectors follows a similar trajectory, while achieving higher precision at higher cutoff values.

On the center plot of Figure 4, we evaluate recall for both classifiers. At very low cutoff values, essentially all permutations are recommended, resulting in high scores (and many FP, hence the low precision values for this cutoff value). The result for the NPMI-classifier is lower as it is based on combinations of terms which are actually seen in literature, while for the word vector approach, since every word is represented as a vector, it is possible to calculate the cosine similarity between all permutation terms. Therefore, in only looking at the left side of the graph, it is difficult to say that the word vector-based classifier is better, although this does seem to be the case when examining the results with higher cut-off values.
The results for the precision and recall show some of the difficulties involved in characterizing the performance of these classifiers. Another way to evaluate these is to create a receiver operating characteristic (ROC) curve (Fawcett, 2004) which plots, for different discrimination levels of a classifier, the true positive rate (TPR) versus the false positive rate (FPR) (Equations (7) and (8); Figure 5).

Classifiers with points strongly in the upper left of the plot perform better than those with their points close to the diagonal. Points on the diagonal indicate that the recommendations are no better than random guessing, while points in the bottom left can be thought of as giving more bad recommendations than good recommendations. Again, the curve for the NPMI-based classifier stops when the discrimination threshold results in zero TP and FP. As one sets the discrimination level for a classifier higher, it often results in allowing predictions of more TP, at the expense of also allowing in more FP. The ROC curve allows for examining this tradeoff while also informing which discrimination level would be useful to prevent the suggested results from being overwhelmed with FP:

\[
TPR = \frac{TP}{TP + FN} \tag{7}
\]

\[
FPR = \frac{FP}{FP + TN} \tag{8}
\]

ROC curves for individual EIPs are included in section S.4 of the Supporting Information and show how well these classifiers predict the actual known linkages at particular sites. What is interesting to note is that for some EIPs (Brownsville), the classifier has very good performance, while for others (Finnish forest industry, Kalundborg), the performance is very poor and the points lie close to, or even below, the diagonal.

The final measure we show is the F-Measure (Equation (9)), which provides a weighted average of the precision and recall, in order to provide a single indicator of the accuracy of the classifiers (Powers, 2007). In the right plot of Figure 4, we see that the word vector-based classifier achieves a roughly 50% higher score than that of the NPMI-based classifier:

\[
F\text{-Measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{9}
\]

6 | DISCUSSION

From the initial proof-of-concept work, several potential benefits and drawbacks emerge for those looking to integrate vector-based analysis into data-centered tools for IS facilitation. We begin our discussion by outlining a few of the more prominent benefits and drawbacks of this approach. This is followed by a look toward some opportunities for further developments in the area, as well as a reflection on the context of such tools.

6.1 | Benefits

One of the strengths showing the most promise is that vector-based approaches can highlight non-explicit combinations in the text corpus, as has been shown with the results comparing matchmaking with word vectors vs. NPMI.
An additional strength is that tools such as fastText, once they have been trained on a large text corpus, can create new word vectors for terms that may not have been seen in the original text. This is because the model creates vectors using what they call “sub-word” information, or aspects related to the substrings found in words. The implications of this are quite powerful as we can very quickly add new terms to a recommendation system, which is something that other approaches struggle with. With other systems, adding new terms can be quite laborious as one has to track down the relevant information needed to provide a useful encoding for the system. This also means that users are not bound by taxonomies. For example, users and developers would not be required to create custom terminology lists or use pre-defined classification lists (e.g., EWC, CPA product classification, industrial classifications, etc), which often lack the detailed definition needed.

Another benefit of this approach is that instead of searching one material at a time, regional analysis on industrial ecosystems with a multitude of organizations and material flows can be performed. This is similar to what was shown in the validation section using the data on existing EIPs, allowing a user to scan for a large range of possible permutations. This can be constructed very quickly and involves first creating vectors for the acceptor and transaction terms, and then calculating the pairwise cosine similarity between the vectors in each set, resulting in a matrix such as that on the right of Figure 3.

Furthermore, as with other machine learning based approaches, this method can integrate new literature as it becomes available, allowing for simplified additional runs on the entire collection of text.

## 6.2 | Drawbacks

One drawback of the approach is that one must still disambiguate the relations highlighted in the results. As such, this method should be augmented with techniques that can relocate the exact source literature analyzed mentioning feedstocks and their corresponding features and properties.

Another significant drawback of this approach is that it does not encode information about the nature of the relation between terms (e.g., why lipids land close to biodiesel). Moreover, no indication of technology readiness level, economic value, or other areas of interest such as GHG savings is encoded in the output of the method.

Another drawback that merits evaluation is that the results of the vector-based approach are not as promising across all verified case studies. For example, this approach misses a lot of potential in the Kalundborg case (see the Supporting Information).

## 6.3 | Potentials for further improvement in machine learning approaches

There are several types of improvements that could be made to the demonstrated approach. First, better named entity recognition could be performed in pre-processing the input text. While our current method relies on part-of-speech tagging, there are more sophisticated options available via services such as DBpedia Spotlight (Mendes et al., 2011). An added benefit is that it is able to recognize synonyms of chemicals and biological names and provide a single standardized term for each. Furthermore, the DBPedia entities, which are identified in the source text function as unique identifiers, are used in several ontologies, and this gives us access to structured data from other databases about these entities.

Second, better models could be used. Recently, BERT (Devlin et al., 2019) and GPT-3 (Brown et al., 2020) have demonstrated significant improvements on NLP tasks. However, some of these models tend to be trained on expensive high-end hardware that may be out of reach for many IS practitioners. One should also consider if these models allow for the creation of vectors for out-of-vocabulary terms.

Third, while we used fastText to train a language model from scratch, it is possible in some cases to download a pre-trained model, and then refine this model further using your own data.

Fourth, it is possible to use word vectors as part of a hybrid system. While the approach demonstrated can be valuable for an initial scan through a large permutation space of acceptors and transactions, we still need to elucidate the explicit relations behind the latent relations. The work by Kuczenski et al. (2016) shows how word vectors can help identify similar LCI processes, and that by van Capelleveen et al. (2021) shows how they can link user-generated terms to EWC codes, illustrate how word vectors can be used to link to identifiers and descriptions in existing databases. This would help in automatically searching for actual known relations and properties which would help explain the high similarity between the examined word vectors. More discussion of hybrid systems is given in Section S.7.

Furthermore, word vectors can be used for numerous tasks. While we highlighted the ability to compare sets of acceptors and transactions, we could repeat this with any set of arbitrary terms. As demonstrated in Section S.6, one can compare terms for waste streams with another set of terms describing specific chemicals, properties, or concerns (e.g., “odor,” “high_tensile_strength,” “combustible,” “hazardous,” “calcium_oxide”).

Taking a broader perspective on data-centered tools for IS facilitation in general, the selection and prioritization of development opportunities will depend on the intended use and context of new and adapted approaches. For those interested in finding low-hanging fruit with high potential impact, integrating additional structured case databases (e.g., on GHG savings, material savings, location and distances within industrial geographies) into a hybrid approach could be a useful hybrid development route.
Regarding the evaluation of such systems, a comparison of the approach to NPMI has been done; however, it would be very interesting to compare and benchmark a broader range of approaches such as the hybrid approach of Gatzioura et al. (2019) and the more traditional approaches listed in Table 1 centering on name or classification matches. Establishing an unbiased method to benchmark approaches with verification enabled through comparison to real-world cases would improve our understanding of the strengths and drawbacks of various existing and novel approaches in the future.

6.4 | Context of data-based approaches

While data-based approaches hold a certain allure due to their anticipated potential for quickly identifying opportunities for resource effectivity initiatives on several geographical scales, the context and limitations of such tools should not be glossed over. The tools the authors have laid out in this article as well as on the isdata GitHub (ISDATA, 2021) are a few building blocks for those interested in constructing and experimenting with their own end-use applications. These building blocks can be assembled and integrated with a wide range of additional information sources in various ways. However, currently, it is not purported that such approaches can out-compete or replace the effectivity of other activities in the facilitation sphere. We hope they would compliment other key facilitation activities such as partnership and trust building, development activities such as research and demonstrating, or new business model development. One example of data-centered and in-person facilitation activities working hand-in-hand is that of International Synergies’ “SYNERGie®4.0” process as summarized on their webpage (International Synergies, 2021).

Surely, the anticipated goal and effect of such undertakings should guide the evaluation of options available and the eventual construction of novel methods. For example, if one has a single material flow for which to examine downstream potentials, alternative approaches such as technology scouting, R&D (internal or outsourced), or simple web-based searches may prove more fitting and effective in the early stages of development.

7 | CONCLUSIONS

The work outlined in this article looks to explore how one could automatically generate word vectors representing waste and feedstocks to elicit similarity (and ultimately substitution potential) between resources based on a broad array of characteristics embedded in the literature. Although precise validation is difficult, through testing on verified case data, it is shown that a vector-based approach shows promise in eliciting more relatedness in terms than a simple co-occurrence (NPMI) text analysis approach. However, in a few cases, the vector-based approach is less effective at identifying proven symbiosis connections. Major benefits of the approach include the ability to analyze substitution potential outside the restraints of limited terminology catalogs. Drawbacks include the current difficulty in tracing back why certain suggestions are made, as well as the lack of structured numerical data (e.g., value, impact) for weighting various options. In combination with additional data, and perhaps integrating other data-based approaches, such tools show promise in assisting in the early phases of facilitation activities, lending insight to potentials that could further be driven through, for example, demonstrations, R&D engineering evaluations, or economic and environmental analysis.

DATA AVAILABILITY STATEMENT

The corpus data that support the findings of this study are available from AcclaimIP (patents corpus) and the Elsevier’s Text Mining API (scientific articles). Restrictions apply to the availability of these data, which were used under license for this study. Data are available at https://www.acclaimip.com/ and https://dev.elsevier.com/ respectively with the access granted from Acclaim and Elsevier. The derived data used to produce the figures of this study are openly available in https://github.com/isdata-org/Industrial-Symbiosis-Word-Vectors.

CONFLICT OF INTEREST
The authors declare no conflict of interest.

ORCID

Chris Davis https://orcid.org/0000-0003-3645-3063
Graham Aid https://orcid.org/0000-0001-6262-0609

REFERENCES

Aid, G., Brandt, N., Lysenkova, M., & Smedberg, N. (2015). Looplocal: A heuristic visualization tool to support the strategic facilitation of industrial symbiosis. Journal of Cleaner Production, 98, 328–335.
Álvarez, R., & Ruiz-Puente, C. (2016). Development of the tool SymbioSys to support the transition towards a circular economy based on industrial symbiosis strategies. Waste and Biomass Valorization, 8(5), 1–10.
Asgari, E., & Mofrad, M. R. K. (2015). Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS ONE, 10(11), 1–15.

Baas, L. (2010). Östergötland: Towards a sustainable region on the basis of industrial symbiosis and renewable energy. In Proceeding of the 16th Industrial Sustainable Development Research Conference 30 May - 1 June, Hong Kong, China.

Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. Proceedings of German Society for Computational Linguistics (GSCL) 2009, 31–40.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. arXiv:2005.14165 [cs], arXiv:2005.14165.

Buckland, M., & Gey, F. (1994). The relationship between recall and precision. Journal of the American Society for Information Science, 45(1), 12–19.

Chertow, M., & Park, J. (2016). Scholarship and practice in industrial symbiosis: 1989–2014, (pp. 87–116). Springer International Publishing.

Costa, L., & Ferrão, P. (2010a). Crossroads between resource recovery and industrial symbiosis networks: Evidences from case studies. Regional Development Dialogue, 31, 1–18.

Costa, L., & Ferrão, P. C. (2010b). A case study of industrial symbiosis development using a middle-out approach. Journal of Cleaner Production, 18(10-11), 984–992.

Cutaia, L., Luciano, A., Barberio, G., Shaffoni, S., Mancuso, E., Scagliarino, C., & La Monica, M. (2015). The experience of the first industrial symbiosis platform in Italy. Environmental Engineering and Management Journal, 14(7), 1521–1533.

Davis, C., & Aid, G. (2021, August 2). Word vectors matching transaction names in EIP Case Studies. https://github.com/isdata-org/Industrial-Symbiosis-Word-Vectors

Davis, C. B., Aid, G., & Zhu, B. (2017). Secondary resources in the bio-based economy: A computer assisted survey of value pathways in academic literature. Waste and Biomass Valorisation, 8(7), 2229–2246.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 [cs], arXiv:1810.04805.

Ding, J., & Hua, W. (2012). Featured chemical industrial parks in China: History, current status and outlook. Resources, Conservation and Recycling, 63, 43–53.

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. Pattern Recognition Letters, 31(8), 1–38.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930–1955 (pp. 1–32). Basil Blackwell.

Gatzioura, A., Sànchez-Marrè, M., & Gibert, K. (2019). A hybrid recommender system to improve circular economy in industrial symbiotic networks. Energies, 12(18), 1–24.

Ghalli, M. R., & Frayret, J. M. (2019). Social semantic web framework for industrial synergies initiation. Journal of Industrial Ecology, 23(3), 726–738.

Ghosh, S., Chakraborty, P., Cohn, E., Brownstein, J. S., & Ramakrishnan, N. (2016). Characterizing diseases from unstructured text: A vocabulary driven word2vec approach. In Proceedings of the 25th ACM International Conference on Information and Knowledge Management, (pp. 1129–1138). ACM Press.

Grant, G. B., Seager, T. P., Massard, G., & Nies, L. (2010). Information and communication technology for industrial symbiosis. Journal of Industrial Ecology, 14(5), 740–753.

Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146–162.

Hein, A., Yannou, B., Jankovic, M., & Farel, R. (2017). Towards an automatized generation of rule-based systems for architecting eco-industrial parks. In: Chakrabarti, A., Chakrabarti, D. (Eds.), ICORD 2017: Research into design for communities (vol. 1). Smart Innovation, Systems and Technologies. Springer. https://link.springer.com/chapter/10.1007%2F978-981-10-3518-0_60

International Synergies (2021). Synergie(R)4.0. https://www.international-synergies.com/what-we-do/synergie40/

ISDATA (2021). ISDATA Github Repository. https://github.com/isdata-org/

Kalundborg Symbiose (2021). Kalundborg Symbiose. http://www.symbiosisk.dk/en

Kuczynski, B., Davis, C. B., Rivela, B., & Janowicz, K. (2016). Semantic catalogs for life cycle assessment data. Journal of Cleaner Production, 137, 1109–1117.

Kusner, M. J., Sun, Y., Kolkin, N. I., & Weinberger, K. Q. (2015). From document to document distances. Proceedings of The 32nd International Conference on Machine Learning, 37, 957–966.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In Advances in neural information processing systems (pp. 2177–2185). Curran Associates.

Li, Y. (2014). Network data and functions for eco-industrial parks. https://github.com/complexly/EIPs

Lombardi, D. R., & Laybourn, P. (2012). Redefining industrial symbiosis: Crossing academic–practitioner boundaries. Journal of Industrial Ecology, 16(1), 28–37.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. Behavior Research Methods, Instruments, & Computers, 28(2), 203–208.

Martin, S., Weitz, K., Cushman, R., & Sharma, A. (1996). Eco-industrial parks: A case study and analysis of economic, environmental, technical, and regulatory issues [Technical report]. Office of Policy, Planning and Evaluation, USEPA, Washington, DC, USA.

Massard, G., & Erkman, S. (2009). A web-GIS tool for industrial symbiosis: Preliminary results and perspectives. 23rd International Conference on Informatics.

Mendes, P. N., Jakob, M., García-Silva, A., & Bizer, C. (2011). DBpedia spotlight: shedding light on the web of documents. In Proceedings of the 7th international conference on semantic systems - i-Semantics '11 (pp. 1–8), Graz, Austria. ACM Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Arxiv, 1–12.

Minarro-Gimenez, J. A., Marin-Alonso, O., & Samwald, M. (2014). Exploring the application of deep learning techniques on medical text corpora. Studies in Health Technology and Informatics, 205, 584–588.

Neves, A., Godina, R., Azevedo, S. G., & Matias, J. C. (2020). A comprehensive review of industrial symbiosis. Journal of Cleaner Production, 247, 119113.

Pakarinen, S., Mattila, T., Melanen, M., & Nissinen, A. (2010). Sustainability and industrial symbiosis: The evolution of a Finnish forest industry complex. Resources, Conservation and Recycling, 54(12), 1393–1404.

Park, H., Rene, E., Choi, S., & Chiu, A. (2008). Strategies for sustainable development of industrial park in Ulsan, South Korea: From spontaneous evolution to systematic expansion of industrial symbiosis. Journal of Environmental Management, 87(1), 1–13.

Park, H.-S., & Behera, S. K. (2014). Methodological aspects of applying eco-efficiency indicators to industrial symbiosis networks. Journal of Cleaner Production, 64, 478–485.
Pecina, P. (2005). An extensive empirical study of collocation extraction methods. In Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (pp. 13–18). Association for Computational Linguistics.

Powers, D. M. W. (2007). Evaluation: From precision, recall and F-Measure to ROC, informedness, markedness and correlation [Technical report]. School of Informatics and Engineering, Flinders University, Adelaide, Australia.

Schwarz, E. J., & Steininger, K. W. (1997). Implementing nature’s lesson: The industrial recycling network enhancing regional development. Journal of Cleaner Production Cleaner Production, 5(12), 47–56.

Shi, H., Chertow, M., & Song, Y. (2010). Developing country experience with eco-industrial parks: A case study of the Tianjin Economic-Technological Development Area in China. Journal of Cleaner Production, 18(3), 191–199.

Simonite, T. (2017). Facebook’s AI chief: Machines could learn common sense from video. Technology Review. https://www.technologyreview.com/2017/03/09/153343/facebook-s-ai-chief-machines-could-learn-common-sense-from-video/

Sinclair Knight Merz (2007). Kwinana industrial area integrated assessment: Environmental, social and economic impact [Technical report], Kwinana Industries Council, Perth, Australia.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding. In Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (pp. 1555–1565). Baltimore, Maryland, USA. Association for Computational Linguistics.

Tang, D., Wei, F., Yang, N., Zhou, M., Liu, T., & Qin, B. (2014). Learning sentiment-specific word embedding. In Proceedings of the 52nd annual meeting of the Association for Computational Linguistics (pp. 1555–1565). Baltimore, Maryland, USA. Association for Computational Linguistics.

Tansey, W., Lowe, E. W., & Scott, J. G. (2016). Diet2Vec: Multi-scale analysis of massive dietary data. https://arxiv.org/abs/1612.00388v1.

Tiejun, D. (2010). Two quantitative indices for the planning and evaluation of eco-industrial parks. Resources, Conservation and Recycling, 54(7), 442–448.

Trokanas, N., Cecelja, F., & Raafat, T. (2014). Semantic input/output matching for waste processing in industrial symbiosis. Computers and Chemical Engineering, 66, 259–268.

van Capelleveen, G., Amrit, C., & Yazan, D. M. (2018). A literature survey of information systems facilitating the identification of industrial symbiosis. In Otjacques, B., Hitzelberger, P., Naumann, S., & Wohlgemuth, V. (Eds.), From science to society (pp. 155–169). Springer International Publishing.

van Capelleveen, G., Amrit, C., Zijm, H., Yazan, D. M., & Abdi, A. (2021). Toward building recommender systems for the circular economy: Exploring the perils of the European Waste Catalogue. Journal of Environmental Management, 277, 1–15.

Wang, Z., Shi, L., Hu, D., Xu, Y., & Sun, D. (2010). Pursuing sustainable industrial development through the eco-industrial parks. Annals of the New York Academy of Sciences, 1195, E145–E153.

Yang, L., Hu, S.-y., Liang, R.-z., Chen, D.-j., Fang, X., Li, Y.-r., Shen, J.-z., Feng, J.-t., & Kong, L.-q. (2004). The Lubei eco-industrial model of China. Chinese Journal of Process Engineering, 4(5), 467–474.

Yang, S., & Feng, N. (2008). A case study of industrial symbiosis: Nanning Sugar Co., Ltd. in China. Resources, Conservation and Recycling, 52(5), 813–820.

Zhu, Q., Lowe, E. A., Wei, Y.-a., & Barnes, D. (2007). Industrial symbiosis in China: A case study of the Guitang group. Journal of Industrial Ecology, 11(1), 31–42.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher’s website.

How to cite this article: Davis C, Aid G. Machine learning-assisted industrial symbiosis: Testing the ability of word vectors to estimate similarity formaterial substitutions. J Ind Ecol. 2022;26:27–43. https://doi.org/10.1111/jiec.13245