Towards an Automatized Generation of Rule- Based systems for architecting eco-industrial parks

Andreas Makoto Hein, Marija Jankovic, Bernard Yannou

Hein A., Jankovic M., Yannou B., "Towards an Automatized Generation of Rule-Based systems for architecting eco-industrial parks", in *6th International Conference on Research into Design (ICoRD'17)*, Guwahati, India, 2017

Towards an Automatized Generation of Rule-Based systems for architecting eco-industrial parks

Abstract: In this article we present the matchmaking problem in industrial symbiosis where wastes from one company is matched with resources of another company that could be substituted. Identifying potential matches is difficult, as it is based on process-specific knowledge that certain wastes can be used for specific processes. Capturing this knowledge in waste-resource matching rules manually is time-consuming. Therefore, we argue that a Natural Language Processing (NLP)-based approach of semi-automatically extracting rules from domain-specific data sets could be a viable approach to solving this problem. The basic NLP problem to solve is to find similar concepts (synonyms), part-whole relationships (meronyms), and "is a" relationships (hyponyms). Synonyms are important for finding wastes and resources that are named differently but refer to the object. Meronyms are part-whole relationships that may help to identify wastes with components that could be used as a resource. Hyponyms allow for building taxonomies. We present the results of an initial literature survey of algorithms that are able to find these relationships in large sets of unstructured text documents. Furthermore, we propose a research approach for further extending the literature survey and testing the existing algorithms on small test cases and realistic matchmaking case. For future work, additional problems that fall into the NLP category can be addressed such as semi-automatically identifying processes for converting wastes into resources.

Introduction

An eco-industrial park is a set of companies within an industrial zone that share resources and thereby increase economic profitability and decrease environmental impact. One of the key underlying concepts of eco-industrial parks is "industrial symbiosis". An industrial symbiosis is the use of an underutilized resource from one actor, such as wastes, as a substitute for a new

resource of another actor. The concept is therefore related to waste recycling. Fig. 1 shows an example of an industrial symbiosis from the eco-industrial park in Kalundborg, Denmark. Instead of using fresh water from a local lake, the coal power plant uses waste water from a close-by oil refinery. The coal power plant saves money, as fresh water is expensive and the oil refinery can save the capital cost for constructing a waste water treatment plant.

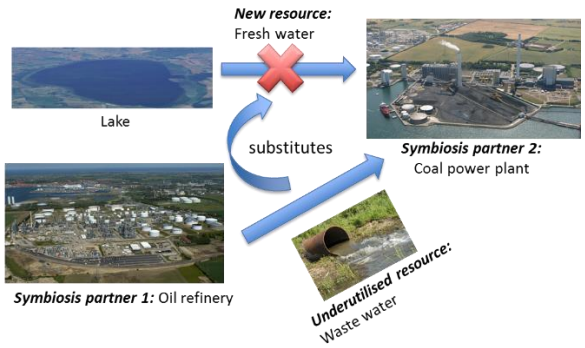

Fig. 1: Example of an industrial symbiosis at the eco-industrial park of Kalundborg [1]

An eco-industrial park is commonly based on a whole network of industrial symbioses. One of the challenges of creating industrial symbioses is that it is knowledge-intensive. For identifying symbiosis opportunities, companies need to share data in order to see if they can make use of another company's underutilized resources. However, companies usually do not share data if they do not see a benefit in sharing a priori. Without sharing data, it is difficult to find symbiosis opportunities.

For solving this problem, we previously developed a prototype of a rule-based system that captured some of the rules and heuristics for identifying symbiosis opportunities without the need of collecting proprietary company data and only using publicly available data. The system is intended to automatically identify promising industrial symbiosis opportunities and generate eco-industrial park architectures that may have a good economic and environmental performance. Similar rule-based systems for Earth observation satellite architectures have recently been developed and presented in [2], [3].

One of the drawbacks of such a static rule-based system is its maintenance and updating [4][5][6]. As new technologies and processes for implementing industrial symbioses are constantly developed, it is important to keep the rule-base up to date. Currently, this is done manually, which is time-consuming and expensive, as the information about the technologies and processes has to be "mined" via extensive literature surveys and expert interviews. However, recent developments in natural language processing (NLP) may have the potential to significantly improve the process of updating rule-bases, underlying taxonomies, and ontologies.

In this paper, we first provide an overview of areas where NLP could add value to rule-based systems.

We start with a literature survey, covering the current state of the art in finding waste – resource pairs in industrial symbiosis and subsequently providing an overview of existing NLP and big data analytics approaches.

Problem formulation

The main objective of the rule-based matchmaking system, in the following called "symbiosis explorer", is to identify opportunities where wastes from one company can be used as a resource by another. These wastes would otherwise be discarded and the other company would instead consume fresh resources [7]. Such waste – resource transfers are called "industrial symbiosis" in the following, as defined by [7]. Industrial symbiosis is attractive in the context of sustainability as it promises to reduce environmental impact and is also feasible economically. However, there are several challenges associated with creating industrial symbioses. One of these challenges is the identification of potential symbiosis opportunities. Identifying these opportunities is not trivial, as there are several obstacles:

- Waste and resource stream data of companies are often confidential.
- Finding opportunities where wastes can substitute resources is not trivial, as it requires technical knowledge about the underlying processes. Furthermore, even when the composition of the waste

and the resource is identical, there are cases where opportunities cannot easily be identified as they are named differently.

One of the shortcomings of existing matchmaking tools is that they require direct input from industrial companies, which presupposes that there is an initial motivation to share data. Another shortcoming is that existing matchmaking systems match wastes and resources based on term identity. It means that if one company creates "plastic" as a waste and another company needs "plastic" as a resource, a match is identified. However, if instead of "plastic" the company creates "plastic bottles", no match is identified. The e-Symbiosis project [8] aims at improving the current state of the art by adding waste and resource taxonomies and allowing for semantic matching. They use a taxonomic distance metric in order to quantify the likelihood that a waste can be matched with a resource.

In our research, we combined several approaches that were already introduced in the literature but are extending these approaches. We propose a system where no input from companies is needed for identifying symbiosis opportunities. Instead, we use so-called meta-models of industrial plants, where types of plants and their usual inputs and outputs are described. It allows for an a priori identification of symbiosis opportunities that can be later refined by more accurate data from the companies. Furthermore, as in the e-Symbiosis tool, we use taxonomies and a set of knowledge-based rules that describe when a certain waste can substitute for a resource.

Our proposal is to improve on this current system by introducing approaches to automatize the generation of rules in order to allow for an automated or at least semi-automated approach to creating matchmaking rules. Such a system would not only be interesting for the area of industrial symbiosis but for recycling markets in general. Fig. 2 shows the architecture of the proposed matchmaking system. From publicly accessible data bases such as books, patent databases, and Wikipedia, synonyms for wastes and resources are identified. In particular, we are interested in:

- Waste and resource compositions that are often introduced in online dictionaries and Wikipedia.
- Identify synonymous terms, e.g. lime is equivalent to calcium oxide or calcium hydroxide.

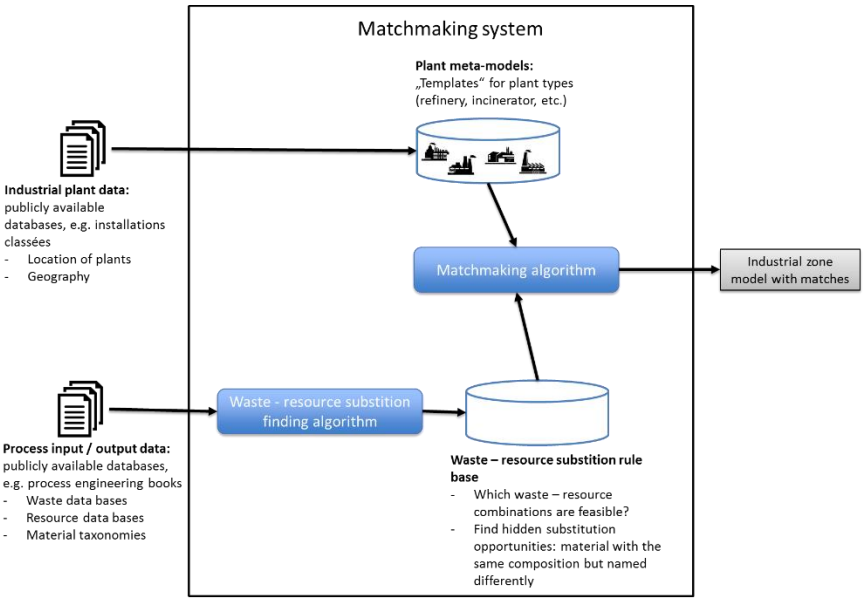

Fig. 2: Architecture of the matchmaking system

Literature survey

Natural Language Processing (NLP) "employs computational techniques for the purpose of learning, understanding, and producing human language content." [9] NLP is of interest for us, as finding synonyms and related terms in a corpus of literature is one of the problems for which NLP approaches exist. In the following, we provide a quick overview of the NLP literature in order to identify approaches that are suitable for addressing our before-stated problem. [9], [10] provide overviews of current NLP approaches and outline possible future trends.

First approaches to NLP date back to the 1950s. First NLP approaches were mostly based on hand-coded rules for automated translation etc. Recent approaches rely on machine learning, mostly grounded in statistical inference-making using large bodies of annotated text. The statistical approaches seem to be better suited for capturing the highly contextual, fuzzy, ambiguous, and dynamic nature of natural languages.

According to [9] NLP can be defined as "the subfield of computer science concerned with using computational techniques to learn, understand, and produce human language content." An up-to-date overview of NLP approaches is presented in [9], [10].

The general problem we are dealing with is the automatic or semi-automatic identification of ontological (lexical) relationships. More specifically we are interested in identifying waste or resources that have different names but have the same meaning (synonyms), part-whole relationships (meronyms) for matching components of a waste or resource, and "is a" relationships (hyponyms) that can be used for building up taxonomies.

Methodologies for creating ontologies from text have been proposed such as in [11]–[15]. An overview of algorithms for identifying synonyms from large bodies of text is shown in Table 1. Latent semantic analysis (LSA) is based on distributional similarity, which is based on the "distributional hypothesis" [16]. The key idea of the distributional hypothesis is that words with similar meaning tend to appear in the similar contexts [17]. However, LSA has shortcomings. As it uses a least-square fitting, it is based on the assumption that language data is normally distributed. Normal distribution is a precondition for least-square fitting. However, language data is not normally distributed [18]. LSA is a purely statistic method. By contrast, Word2Vec is a machine learning approach, based on a 2-layered neural network [19], [20]. The neural network is trained on a text corpus to guess the words that appear in its context. Based on the distributional hypothesis, words with similar meaning tend to appear in similar contexts. Hence, Word2Vec can be used for comparing contextual words of different concepts. The more similar the contextual words, the more likely the concepts have the same meaning. As Word2Vec is based on a neural network, extensive training data is required. A third approach, called DFEAT, is based on supervised learning and distributional features [21]. The value of the distributional feature indicates the commonality of the context of a word pair. Using a pattern-matching algorithm, the algorithm is trained with a test set of words. Compared to other synonym classifiers on the test set, the DFEAT algorithm showed a superior performance.

Table 1: NLP algorithms for detecting synonyms

NLP approach	Input	Algorithm	Output	Key references
Latent semantic analysis	Documents	Singular Value Decomposition	Term set from document with reduced dimensionality	[22], [23]
Word2Vec	Candidate concept	2-layered neural network	Ranked list of synonyms	[24]
DFEAT	Training documents and application documents	Supervised learning and distributional features	Ranked list of synonyms	[21]

Similar to finding synonyms, a number of algorithms have been proposed for identifying part-whole relationships. [25] presents an approach for finding part-whole relationships in domain-specific data. Part-whole relationships in domain-specific data is more difficult to address than such relationships in general language, as the available data is much smaller. As a consequence, it is more difficult to train machine learning algorithms. The approach proposed in [25] leverages on part-whole relationships extracted from an open-domain corpus and extends these relationships by domain-specific relationships. [26] is arguably the first automatic part extraction from a large, unlabeled, corpus with a reported precision of 0.55. The purpose of the method is to add the discovered part-whole relationships to an existing ontology or in a semantic lexicon. The algorithm rank-orders words that are candidates for parts of a whole, e.g. "speedometer" as part of a "car". [27] presents an algorithm for identifying meronyms in biomedical text. They use the PartEx unsupervised learning algorithm that learns part-whole patterns from biomedical knowledge bases and infers part-whole relationships in yet unknown data. In its reported version it achieves a recall of 0.73 and a precision of 0.58. The author claims that no manual labelling of the corpus and manual selection of patterns is needed.

Regarding hyponyms, [28] presents an algorithm for identifying hyponyms in large corpora. [29] presents a semi-automatic approach for creating taxonomies from text on websites in the context of the semantic web. [30] presents a method for extracting ontologies from semi-structured information and text documents in domain-specific corporate intranets.

Integrating these algorithms with other powerful NLP algorithms such as parsers has become easier due to the availability of several open source NLP tools such as GATE NLP, Stanford Core NLP Suite, Apache OpenNLP, and the Natural Language Toolkit. Furthermore, the availability of extensive material, resources, and waste taxonomies provide the data on which the algorithms can be run.

Proposed Research Approach

Based on the initial literature survey presented in the previous section, we propose a research approach for coming up with a matchmaking approach that is based on an automated or semi-automated data base of waste – resource substitution relationships. Fig. 3 depicts the steps of the research approach. First, a more extended literature survey on existing algorithms for finding synonyms, meronyms, and hyponyms is conducted. In case these algorithms are available in open source mode or can be rapidly implemented, they may be tested on small data sets for verifying their applicability to the matchmaking problem. The test results are compared and the most adequate algorithms in terms of applicability and efficiency selected. This subset of algorithms is then applied to a more realistic data set with waste – resource data. The resulting waste – resource pairs are then validated by experts. On the basis of the validation the algorithms precision with respect to the data set can be calculated.

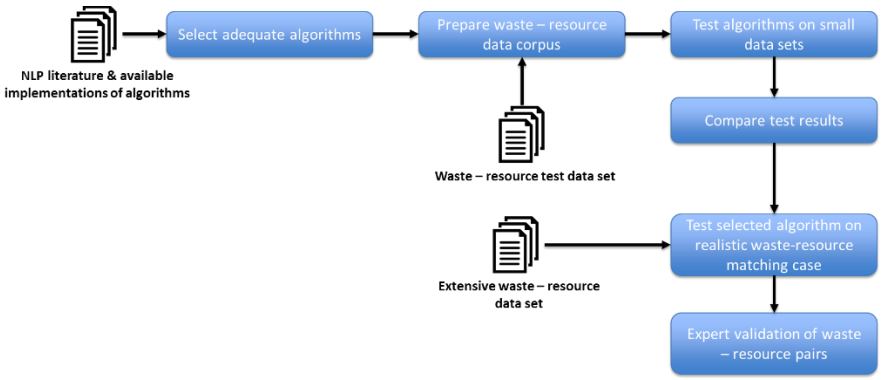

Fig. 3: Research approach for finding and testing matchmaking algorithms

Conclusions

In this article we presented the matchmaking problem in industrial symbiosis. Identifying potential matches is difficult, as it is based on process-specific knowledge that certain wastes can be used for specific processes. However, capturing this knowledge in waste-resource matching rules manually is time-consuming. Therefore, we propose a NLP-based approach of semi-automatically extracting ontological relationships from domain-specific data sets that can be used as a basis for matching rules. We presented the results of an initial literature survey of algorithms that are able to find ontological relationships in large sets of unstructured text documents. Furthermore, we proposed a research approach for further extending the literature survey and to test the existing algorithms on small test cases and a realistic matchmaking case. For future work, additional problems that fall into the NLP category can be addressed such as semi-automatically identifying processes for converting wastes to resources.

References

[1] A. Hein, M. Jankovic, R. Farel, and B. Yannou, "A DATA- AND KNOWLEDGE-DRIVEN METHODOLOGY FOR GENERATING ECO-INDUSTRIAL PARK ARCHITECTURES," in *Proceedings of the ASME 2016 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference IDETC/CIE 2016*, 2016.

[2] D. Selva Valero, "Rule-based system architecting of Earth observation satellite systems," Massachusetts Institute of Technology, 2012.

[3] S. Das, D. Selva, and A. Golkar, "An Intelligent Spacecraft Configuration Tool for Mission Architecture Space Exploration," *AIAA Sp. 2015 Conf. …*, 2015.

[4] X. Li, "Quality time-What's so bad about rule-based programming?," *Software, IEEE*, 1991.

[5] J. Vargas and S. Raj, "Developing maintainable expert systems using case- based reasoning," *Expert Syst.*, 1993.

[6] E. Sollow, "Assessing the Maintainability of XCQN-in-RIME: Coping with the Problems of a VERY Large Rule-Base," 1987.

[7] P. Deutz, "Food for thought: Seeking the essence of industrial symbiosis," *Pathways to Environ. Sustain.*, 2014.

[8] F. Cecelja, T. Raafat, and N. Trokanas, "e-Symbiosis: technology-enabled support for Industrial Symbiosis targeting Small and Medium Enterprises and innovation," *J. Clean. …*, 2015.

[9] J. Hirschberg and C. Manning, “Advances in natural language processing,” *Science (80-.).*, 2015.

[10] E. Cambria and B. White, “Jumping NLP curves: a review of natural language processing research,” *Comput. Intell. Mag. …*, 2014.

[11] A. Maedche and S. Staab, “Semi-automatic engineering of ontologies from text,” *Proc. 12th Int. Conf. …*, 2000.

[12] T. Wächter and M. Schroeder, “Semi-automated ontology generation within OBO-Edit,” *Bioinformatics*, 2010.

[13] C. Lee, Y. Kao, Y. Kuo, and M. Wang, “Automated ontology construction for unstructured text documents,” *Data Knowl. Eng.*, 2007.

[14] P. Gawrysiak and G. Protaziuk, “Text onto miner–A semi automated ontology building system,” *Found. Intell. …*, 2008.

[15] B. Fortuna, M. Grobelnik, and D. Mladenic, “Semi-automatic data-driven ontology construction system,” *Proc. 9th Int. …*, 2006.

[16] Z. Harris, “Distributional structure,” *Word*, 1954.

[17] J. Firth, “{A synopsis of linguistic theory, 1930-1955},” 1957.

[18] T. Van de Cruys, “Distributional Similarity - Overview and Applications,” 2010.

[19] Y. Goldberg and O. Levy, “word2vec explained: Deriving mikolov et al.’s negative-sampling word-embedding method,” *arXiv Prepr. arXiv1402.3722*, 2014.

[20] T. Mikolov, I. Sutskever, and K. Chen, “Distributed representations of words and phrases and their compositionality,” *Adv. neural …*, 2013.

[21] M. Hagiwara, “A supervised learning approach to automatic synonym identification based on distributional features,” *Proc. 46th Annu. Meet. …*, 2008.

[22] T. Landauer, P. Foltz, and D. Laham, “An introduction to latent semantic analysis,” *Discourse Process.*, 1998.

[23] C. Papadimitriou and H. Tamaki, “Latent semantic indexing: A probabilistic analysis,” *Proc. …*, 1998.

[24] C. Wang, L. Cao, and B. Zhou, “Medical synonym extraction with concept space models,” *arXiv Prepr. arXiv1506.00528*, 2015.

[25] A. Ittoo, G. Bouma, L. Maruster, and H. Wortmann, “Extracting Meronymy Relationships from Domain-Specific, Textual Corporate Databases,” in *Natural Language Processing and Information Systems Volume 6177 of the series Lecture Notes in Computer Science*, 2010, p. pp 48–59.

[26] M. Berland and E. Charniak, “Finding parts in very large corpora,” *Proc. 37th Annu. Meet. …*, 1999.

[27] A. Roberts, “Learning meronyms from biomedical text,” *Proc. ACL Student Res. Work.*, 2005.

[28] M. Hearst, “Automatic acquisition of hyponyms from large text corpora,” *Proc. 14th Conf. Comput. …*, 1992.

[29] A. Maedche and S. Staab, “Ontology learning,” *Handb. Ontol.*, 2004.

[30] J. Kietz, A. Maedche, and R. Volz, “A method for semi-automatic ontology acquisition from a corporate intranet,” *… Text”, Juan-Les-Pins, Fr. …*, 2000.