Integrating Corpus Statistics into Measures of Business Process Similarity

Completed Research Paper

Marcus Fischer
Julius Maximilians University Würzburg
marcus.fischer@uni-wuerzburg.de

Florian Imgrund
Julius Maximilians University Würzburg
florian.imgrund@uni-wuerzburg.de

Christian Janiesch
Julius Maximilians University Würzburg
christian.janiesch@uni-wuerzburg.de

Axel Winkelmann
Julius Maximilians University Würzburg
axel.winkelmann@uni-wuerzburg.de

Abstract

In a rapidly changing environment, organizations must adapt their business processes continuously. While numerous methods enable enterprises to conceptualize and analyze their organizational structure, the task of business process modeling remains complex and time-consuming. However, by reusing and adapting existing process models, enterprises can reduce the task's complexity while improving the quality of results. To facilitate the identification of adaptable processes, several techniques of business process similarity (BPS) have been proposed in recent years. Although most approaches produce sound results in controlled evaluations, this paper argues that their applicability is limited when analyzing real-world processes, which do not fully comply with notational labeling specifications. Consequently, we aim to enhance existing BPS techniques by using corpus statistics to account for the explanatory power of words within labels of process models. Results from our evaluation suggest that corpus statistics can improve BPS computations and can positively influence the quality of practical implications.

Keywords

Business Process Management, Business Process Similarity, Corpus Statistics, Collection Management

Introduction

Today's enterprises operate in a rapidly changing environment. Due to the ongoing globalization and technological advancements, market pressure increased tremendously in recent years. Thus, enterprises must actively manage and optimize their business processes to ensure their long-term competitiveness. However, to capture potential benefits from Business Process Management (BPM), enterprises must document and model their organizational structure appropriately. While various process modeling techniques can be utilized to construct an analytical foundation for BPM activities, the task itself remains time-consuming and is biased by modelers' individual perceptions (Becker et al. 2000).

As digitization and supply chain integration result in an increasing process complexity, reusing an industry's best practices can reduce BPM-induced costs and improve process quality (Martens et al. 2014). However, syntactic and semantic ambiguities can hamper potential benefits if process modeling is based on an enterprise-specific vocabulary. In fact, if enterprises use external documentations to support their operations, they need to control for synonyms and homonyms in order to avoid misunderstandings (Dijkman et al. 2011). In addition, extensive experience and expert knowledge is necessary to identify and fully harness organizational improvement potentials. Recently, numerous contributions have introduced methods and techniques to match business process models (Dijkman et al. 2009; Dijkman et al. 2011; van Dongen et al. 2013), mostly focusing on the use of distance measures and lexical-semantic databases to identify similar fragments of business processes and their underlying process graphs.

However, these approaches tend to overlook the structure and composition of labels comprised by a process model. Using the example of the Event-Driven Process Chain (EPC), the labeling of functions is formally limited to an object-verb-combination (Hoffmann et al. 1993). To address various organizational demands, these notational specifications are frequently modified. Analyzing our real-world process repository, the average number of words for labeling functions adds up to 5.2. While a more detailed labeling can increase the understandability of process models, we can typically neglect the explanatory power of additional words (Li et al. 2006). However, as existing approaches mostly consider full labels for business process similarity (BPS) computation, results frequently lack efficiency and accuracy. Consequently, we summarize our research question as follows:

"How can we improve the efficiency and accuracy of existing BPS methods when evaluating real-world process data?"

To address this research question, we apply a Design Science Research (DSR) methodology to develop our contribution. As our research progressed, we iteratively developed the configuration of the artifact (Baskerville et al. 2009). Following Gregor and Hevner's knowledge contribution framework, we consider our DSR contribution an exaptation of corpus statistics to measures of process similarity in the field of BPM (Gregor and Hevner 2013). It has explanatory power and provides design practice theory for the design and improvement of further methods that aim to improve BPS techniques.

Following the DSR paradigm, our contribution is organized as follows: Section 2 provides an overview of related work on BPS. Necessary theoretical foundations are introduced in Section 3, and Section 4 presents current methods for BPS computation. Subsequently, Section 5 integrates the concept of corpus statistics into the predefined BPS measures, which are experimentally evaluated in Section 6. Finally, Section 7 concludes with a summary of findings, limitations, and future research potentials.

Related Work

As BPM generally comprises a body of methods, techniques, as well as tools and systems to identify, prioritize, analyze, improve, and monitor business processes, BPS can support several activities within established BPM lifecycle models (Dumas et al. 2013). The relevance of BPS methods tends to increase as more enterprises aim to address dynamically changing market conditions by holistic BPM approaches, based on collaboration and decentralization (Imgrund et al. 2017). As many organizational stakeholders participate in BPM activities, such as process identification and discovery, enterprises can access a large amount of previously neglected business processes that typically exceeds the capabilities of traditional BPM approaches by far. To ensure the manageability of these growing process repositories, methods of BPS provide means to structure and organize processes, e.g. in terms of detecting redundancies, inconsistencies, and notational violations.

As a conceptual foundation, Dijkman et al. (2011) provide a detailed overview of existing methods for computing BPS and their practical evaluation. In fact, BPS techniques are categorized into metrics for comparing single process elements or full-scale process models, with both categories closely related, as the similarity of process models is based on a mapping of their corresponding process elements. While syntactic similarity is determined by the edit-distance of two labels, semantic similarity accounts for synonyms and homonyms by integrating *WordNet* as a semantic-lexical database (Manning and Schütze 1999). Rudimentary evaluating a word's importance for the meaning of a label, frequently occurring words like 'a', 'an', and 'for' are eliminated. The resulting mapping is used to compute the label matching, structural, and behavioral similarity of business processes (Dumas et al. 2009).

Li et al. (2006) measure process similarity as the sum of the high-level operations *insertion*, *elimination*, *substitution,* and *movement* to transform one business process model into another. In addition, Ehrig et al. (2007) introduce an ontology-based semantic modeling approach ensuring process model interoperability and interconnectivity. In fact, they account for the semantic information of each word within a process element's label but do not consider its relevance for the label's meaning. Van Dongen et al. (2013) widen the scope of BPS by evaluating an EPC's functions regarding their expected behavior. After determining syntactic and semantic similarity based on established BPS methods, the mapping is enhanced by contextual information of a function's preceding and succeeding events. In domains that offer a standardized and controlled vocabulary, Akkiraju and Ivan (2010) analyze the similarity of two processes by computing the ratio of syntactical equally labeled activities to the total number of labels.

The Triple-S approach uses the syntactic, semantic, and structural characteristics of process models to compute a combined matching score (Cayoglu et al. 2014). The N-Ary Semantic Cluster Matching technique further proposes a clustering approach that uses a n-ary cluster matcher based on the measure of semantic similarity (Cayoglu et al. 2014). The Extended Semantic Greedy Matching approach utilizes the transition of process models to construct a set of pairwise transition matches (REF). Cayoglu et al. (2014) further compute the similarity of two labels by treating each label as a bag of words. Subsequently, they perform word stemming and use established syntactic and semantic techniques for similarity computation. In addition, the Process Matching Using Positional Language Models approach adopts language modeling methods from the domain of Information Retrieval (Cayoglu et al. 2014). Particularly, it uses passage-based modeling and considers structural features of process models by positional language modeling. As a result, the method creates a similarity matrix between process elements and derives correspondences using second line matching.

As many metrics aim to establish an exact matching between two process representations, they are typically characterized by a high computation time and complexity, considerably hampering their applicability when analyzing and managing large process repositories. To address corresponding challenges, Awad (2007) introduces BPMN-Q, which allows enterprises to query a repository for similar business process models with a *binary match* or no *match* result. In line with that, Yan et al. (2010) abstract business processes to a set of constitutive features. Thus, processes stored within a process repository are evaluated against these features to establish a similarity ranking by applying feature-based similarity. Consequently, syntactical, semantic, and structural characteristics are utilized for process abstraction, while the similarity ranking itself results from the ratio of features matched to the total number of features.

While traditional BPS methods have proven to produce sound results in their various fields of application, they do not adequately account for the meaning of words comprised by a process elements label. In collaborative environments especially, where all stakeholders of an organization can participate in process modeling and typically construct models that do not fully comply with notational specifications, their applicability is limited. To address these challenges, we argue that integrating corpus statistics can improve the quality and reliability of BPS methods.

Theoretical Background

To provide a theoretical foundation for the following sections, relevant preliminaries are presented subsequently. Since our German process repository is modeled in EPC notation, its basic features and a mathematical formalism are introduced at first. The modeling language aims primarily at the description of a process's underlying business logic, than on its formal specification (van Der Aalst 1999). An EPC contains functions, events, and connectors, which are linked by arcs representing the control and information flow. To ensure syntactical correctness, functions and events need to be strictly alternating. While functions represent a task or activity, events can either be start-points, or end-points, or can mark the pre-conditions or post-conditions of functions. As suggested by the academic literature, functions are labeled by a combination of the object to be processed and the necessary activity. Similarly, labels of events formally comprise the processed object and the executed activity (Hoffmann et al. 1993). Eventually, connectors identify logical decision points within a process and can either be conjunctive ("AND"), adjunctive ("OR"), or disjunctive ("XOR") (Thomas and Fellmann 2006).

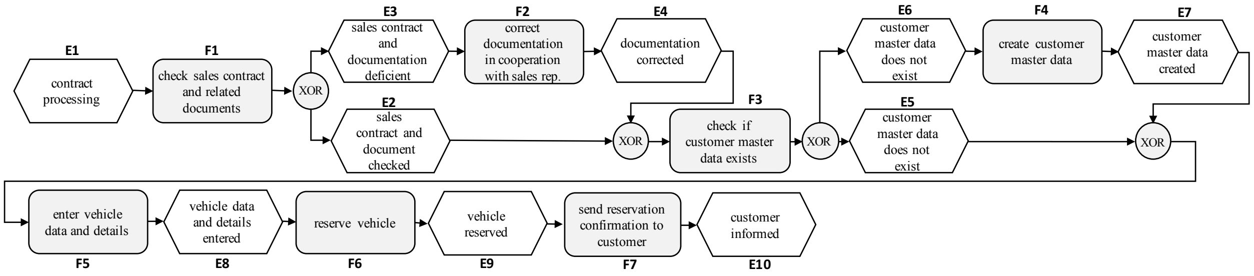
E1
contract processing
F1
check sales contract and related documents
XOR
E3
sales contract and documentation deficient
F2
correct documentation in cooperation with sales rep.
E4
documentation corrected
E2
sales contract and document checked
XOR
F3
check if customer master data exists
XOR
E6
customer master data does not exist
F4
create customer master data
E7
customer master data created
E5
customer master data does not exist
XOR
enter vehicle data and details
F5
vehicle data and details entered
E8
reserve vehicle
F6
vehicle reserved
E9
send reservation confirmation to customer
F7
customer informed
E10

Figure 1: EPC Notation for Business Process Modeling

Figure 1 illustrates the control flow structure of a business process modeled in EPC. The model describes a fragment of the order handling reference process extracted from the ARIS reference model (Scheer and Nüttgens 2000). The process is initiated by the activity of contract registration. The contract is then checked regarding its compliance with formal requirements and is either approved or scheduled for revision. After approval, customer data is handled within a information system and a customer-specific reservation is made eventually. The process terminates after the customer's notification.

We further define an EPC as a directed graph to analyze its structural characteristics and to apply graph-based calculation schemes. Thus, methods presented in this paper can be applied to every other graph-based modeling language like Business Process Modeling and Notation or the Unified Modeling Language (Dijkman et al. 2009). Consequently, Definition 1 introduces a set-theoretic abstraction of business processes.

Definition 1 (Business Process Graph). *A directed process graph is a tuple (N, E, l), in which:*

- *N is a finite set of nodes;*
- *$E \sqsubseteq N \times N$ is a finite set of edges; and*
- *$l: N \rightarrow \varsigma$ is a function that assigns nodes to labels.*

Following Definition 1, an EPC's functions, events, and connectors are replaced by nodes, while arcs are replaced by edges (van Dongen et al. 2013). To analyze full-scale process models, existing metrics to compare process elements can be divided into type-based, syntactic, semantic, and contextual similarity measures. All metrics return a similarity score between 0 and 1, with 0 implying that the compared elements are entirely different and 1 implying their equality (Dijkman et al. 2011).

As similarity computations should be processed automatically, a mechanism must be defined that guarantees the comparison of equal element types (Yan et al. 2010). The formalism presented in Definition 2 ensures that functions are compared with functions and events are compared with events respectively. Subsequently, using *Type* as a precondition for further similarity computations, the matching of different types of elements is avoided.

Definition 1 (Type-based Similarity). *Let EPC_1 and EPC_2 be two disjoint EPCs described by functions, events and connectors. Additionally, let $n \in EPC_1$ and $m \in EPC_2$ be two nodes from the EPCs' underlying process graph. Then 'type' is a binary function that returns 1, if and only if the types of the compared nodes are identical and 0 otherwise.*

$$type\ (n, m) = \begin{cases} 1 & Type\ (n) = Type\ (m) \\ 0 & Type\ (n) \neq Type\ (m) \end{cases}$$

Syntactic equivalence is obtained by adding the number of the edit operations *insertion*, *deletion*, and *substitution* to transform one label into another. To measure syntactic similarity, we use the *Levenshtein* Algorithm (Levenshtein 1966).

Definition 3 (Syntactic Similarity). *Let EPC_1 and EPC_2 be two disjoint EPCs described by $EPC_n = (E_n, F_n, C_n, l_n, A_n)$. Additionally, let $n \in EPC_1$ and $m \in EPC_2$ be two nodes from those EPCs. The nodes are labeled by $l_1(n)$ and $l_2(m)$ with their corresponding length $|l_1(n)|$ and $|l_2(m)|$. Then syntactic similarity can be calculated using the following formalism:*

$$syn(n,m) = type(n,m) \times (1 - \frac{\mathrm{ed}(l_1(n), l_2(m))}{\max(|l_1(n)|, |l_2(m)|)})$$

Although labels are considered as highly representative process features, focusing exclusively on their syntax can result in ambiguity-induced inconsistencies because of synonyms and homonyms (Dijkman et al. 2011). Thus, the similarity of process elements depends strongly on individual perceptions, negatively affecting the quality of results. Consequently, we introduce semantic similarity in Definition 4. For process elements of the same type, semantic similarity is based on weighting synonyms and identical words with specific weighting factors. Dijkman et al. (2011) suggest a weighting factor of 0.75 for synonyms and 1 for identical words, which we set in this paper accordingly. Subsequently, semantic similarity is computed by the sum of weights divided by the number of words contained by the longest label. "Synonyms" represents a binary function that returns 1 if the compared words are synonyms and 0 otherwise.

Definition 4 (Semantic Similarity). *Let EPC_1 and EPC_2 be two disjoint EPCs described by $EPC_n = (E_n, F_n, C_n, l_n, A_n)$. Additionally, let $n \in EPC_1$ and $m \in EPC_2$ be two nodes from those EPCs labeled by the words $w_1 = l_1(n)$ and $w_2 = l_2(m)$ with their corresponding number of words $|w_n|$. Then semantic similarity can be calculated using the following formalism:*

$$sem(n,m) = type(n,m) \times \frac{1.0 \times (w_1 \cap w_2) + 0.75 \times \sum_{\substack{s \in w1/w2 \\ t \in w2/w1}} synonyms\ (s,t)}{\max(|w_1|,|w_2|)}$$

Aiming to detect synonyms, we use the semantic-lexical database *GermaNet* (Hamp and Feldweg 1997). However, to reduce noise within labels, Dijkman et al. (2011) suggest several preparatory steps. First, frequently occurring words like 'a', 'an', and 'for' are eliminated. Second, inflectional and derivationally related words are reduced to their infinitive forms by using the technique of word stemming (Porter 1980). While the introduced techniques exclusively account for the label of each process element, contextual similarity focuses on their preceding and succeeding process elements (van Dongen et al. 2013). In case of an EPC, the similarity score of two functions is determined by the similarity of their input and output events (Dijkman et al. 2011). To account for this contextual information, we establish an equivalence mapping between two process models as introduced in Definition 5.

Definition 5 (Equivalence Mapping). *Let EPC1 and EPC2 be two disjoint EPC's described by $EPC_n = (E_n, F_n, C_n, l_n, A_n)$, and let L_n be their corresponding labels. Additionally, let $l(syn + sem)$: $L_1 \times L_2 \rightarrow [0 \ldots 1]$ be a similarity function that guarantees for all $l_1 \in L_1$ and $l_2 \in L_2$: $lsim(l_1, l_2) = lsim(l_2, l_1)$. An optimal equivalence mapping M^{opt}_{lsim}: $L_1 \nrightarrow L_2$ is an equivalence mapping, such that for all other equivalence mappings M holds that:*

$$\sum\nolimits_{(l_1,l_2) \in M^{opt}_{l(syn+sem)}} lsim(l_1,l_2) \geq \sum\nolimits_{(l_1,l_2) \in M_{l(syn+sem)}} lsim(l_1,l_2)$$

Based on the optimal equivalence mapping of two EPCs, contextual similarity can be computed using the calculation scheme summarized in Definition 6.

Definition 6 (Contextual Similarity). *Let EPC_1 and EPC_2 be two disjoint EPCs described by $EPC_n = (E_n, F_n, C_n, l_n, A_n)$ with $n_n \in F_n \cup E_n$ and let $l_{(syn+sem)}$ be a similarity function. Additionally, let M^{optin}_{lsim}: $n^{in}_1 \nrightarrow n^{in}_2$ and M^{optout}_{lsim}: $n^{out}_1 \nrightarrow n^{out}_2$ be two optimal equivalence mappings between the preceding and succeeding nodes of n_1 and n_2.*

$$con(n,m) = type(n,m) \times \frac{|M^{optin}_{l(syn+sem)}|}{2 \cdot \sqrt{|n^{in}|} \cdot |m^{in}|} + \frac{|M^{optout}_{l(syn+sem)}|}{2 \cdot \sqrt{|n^{out}|} \cdot |m^{out}|}$$

Corpus Statistics for Business Process Similarity Computation

Using traditional BPS methods, we argue that the accuracy of results depends strongly on the length of the compared labels. Thus, accurate results are likely only if process elements labels comply with the notational specifications introduced in the previous sections. By contrast, if labels contain additional information to improve their understandability, traditional techniques show considerable limitations. Thus, this paper proposes a novel approach for BPS computation, which is summarized in Figure 2.

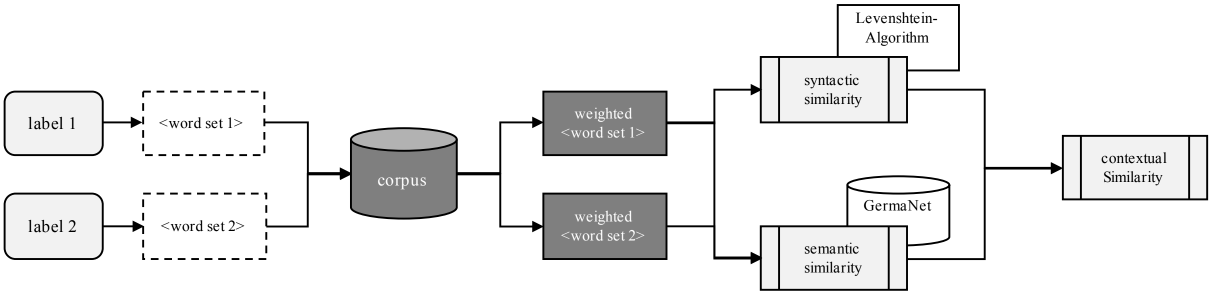

Figure 2: Business Process Similarity Computation Diagram

As the approach is an enhancement of current methods, underlying computation schemes can be modified easily. We define a process element's label as a sequence of words, each carrying useful information. However, the degree to which a word contributes to the meaning of a label can vary significantly. We refer to that degree as a words information content (IR), which we control for by evaluating a label's words against a text corpus. We adopted this procedure from Li et al. (2006), who utilize the concept of IC to determine the similarity of sentences by assigning weighting factors to each word of a sentence according to its IC. Words with a limited explanatory power for the sentence's overall meaning are neglected and thus the accuracy of sentence similarity is improved. As our process repository consists exclusively of EPCs modeled in German language, we use the Leipzig Corpora Collection as a text corpus for our evaluation (Biemann et al. 2007). Nevertheless, the suggested technique is independent of a single natural language. In the case of an English repository, it can easily be performed using the Brown Corpus (Kucera and Francis 1979). The Leipzig Corpora Collection contains over 26 million sentences with more than 425 million tokens and is therefore considered as a representative collection of the German language. Meadow et al. (1992) have shown that words occurring with a lower frequency contain more information than those with a higher frequency (Meadow et al. 1992). Thus, we can utilize this data when determining the IC of a word by evaluating its probability of occurrence within the corpus (Li et al. 2006). For example, the German equivalent of "fast" occurs 96.674 times in total. By contrast, the German equivalent of "customer" only occurs 11.553 times. Thus, according to its IC, the word "customer" has a higher contribution to a label's meaning than "fast". To incorporate this information into our BPS computations, we assign a weighting factor for a word's relative IC compared to all the other words within that corpus. Definition 7 summarizes the proposed formalism.

Definition 7 (Probability of Occurrence). *Let EPC_1 and EPC_2 be two disjoint EPCs described by $EPC_n = (E_n, F_n, C_n, l_n, A_n)$. Additionally, let $n \in EPC_1$ and $m \in EPC_2$ be two nodes from those EPCs labeled by the words $w_1 = l_1(n)$ and $w_2 = l_2(m)$. Ψ defines the probability that a word w_n, occurs within a lingual corpus.*

$$\Psi(w_n) = \frac{n+1}{N+1}$$

According to Definition 7, the probability for a word to occur in a particular corpus results from the ratio of its frequency in the corpus n and its total number of words N (Li et al. 2006). Subsequently, the IC of w_n is computed using Definition 8.

Definition 8 (Weighted Information Content). *Let EPC_1 and EPC_2 be two disjoint EPCs described by $EPC_n = (E_n, F_n, C_n, l_n, A_n)$. Additionally, let $n \in EPC_1$ and $m \in EPC_2$ be two nodes from those EPCs labeled by the words $w_1 = l_1(n)$ and $w_2 = l_2(m)$. Furthermore Ψ refers to a word's probability of occurrence. Then $l(W_n)$ describes its weighted information content.*

$$I(w_n) = 1 - \frac{\log(\mathrm{n}+1)}{\log(N+1)}$$

Experimental Evaluation

After we introduced the measure of IC, the following section aims to evaluate its integration into traditional BPS techniques. Gregor and Hevner (2013) argue that a proof-of-concept is a suitable first step when evaluating DSR.

Implementation Setup

Various authors use reference models for the evaluation of BPS methods (Dijkman et al. 2009; Kunze et al. 2011). Although a commonly accepted definition is not provided in the academic literature, reference models are generally universally applicable, reusable, and contain a set of best practices for an application domain (Thomas 2006). Since reference models are abstracted from a set of individual process models, we selected the "Sales Order Process" from the ARIS Reference Model for our evaluation, which has already been introduced in Figure 1. However, as our n similarity of process elements, we modified each label of this process model while leaving the structure unchanged. Due to this additional information, the process's understandability is increased and a more detailed description of the task is provided. However, while additional information can be useful within an enterprise, it hampers the inter-organizational comparability of process models, e.g. to identify potential synergies resulting from the merger of two enterprises. The modified reference process is illustrated in Figure 3. Consequently, we applied the following modifications:

1. References to the task, that are not specified in other process models, were added.
2. Verbs were attributed by words like "fast", "accurate" or "thoroughly".
3. Unnecessary information such as "via email or letter" were added.

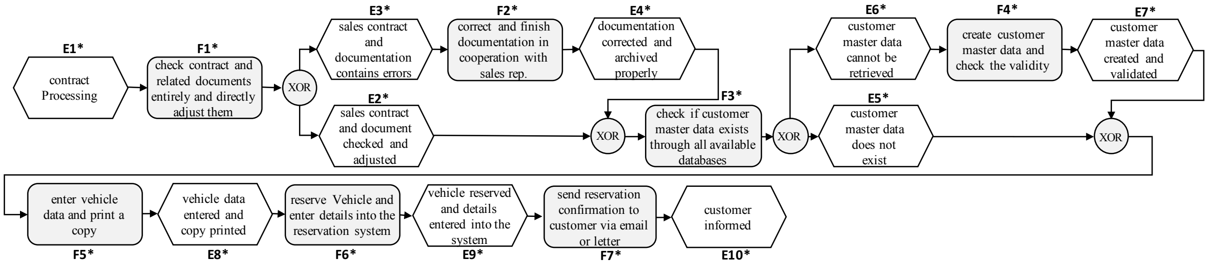
E1*
contract Processing
F1*
check contract and related documents entirely and directly adjust them
XOR
E3*
sales contract and documentation contains errors
F2*
correct and finish documentation in cooperation with sales rep.
E4*
documentation corrected and archived properly
E2*
sales contract and document checked and adjusted
XOR
F3*
check if customer master data exists through all available databases
XOR
E6*
customer master data cannot be retrieved
F4*
create customer master data and check the validity
E7*
customer master data created and validated
E5*
customer master data does not exist
XOR
F5*
enter vehicle data and print a copy
E8*
vehicle data entered and copy printed
F6*
reserve Vehicle and enter details into the reservation system
E9*
vehicle reserved and details entered into the system
F7*
send reservation confirmation to customer via email or letter
E10*
customer informed

Figure 3: Modified Reference Process Model

Analyzing both process models, a total of 17 process elements must be compared. The results we report here use GermaNet for the detection of synonyms and homonyms. However, due to the computational complexity, we reduced the corpus to a fragment of 300.000 randomly selected sentences, containing more than 3 million words. Although we evaluated all 149 element combinations of the same type, results are limited to optimal matches only.

Evaluation

Table 1 summarizes the results of our evaluation. Initially, four experienced process modelers investigated both process variants and established an optimal mapping based on human judgment. Resulting combinations of process elements are presented in the first column. By using the given references, similar process elements can be identified in Figure 1 and 3. Subsequently, the processes were evaluated using syntactic, semantic, and contextual similarity with and without IC enabled. The remaining columns summarize the generated similarity scores.

Table 1: Similarity Measures including Corpus Statistics

human judgment (I)	syntactic similarity (II)	syntactic similarity with IC (III)	semantic similarity (IV)	semantic similarity with IC (V)	contextual similarity (VI)	contextual similarity with IC (VII)
(E1) to (E1*)	0.78	**0.79**	0.50	**0.54**	0.77	**0.78**
(F1) to (F1*)	0.41	**0.54**	0.64	**0.58**	0.82	**0.83**
(E2) to (E2*)	0.55	**0.83**	0.67	**0.57**	1.0	**1.0**
(E3) to (E3*)	0.78	**0.86**	0.60	**0.70**	0.87	**0.9**
(F2) to (F2*)	0.77	**0.85**	0.70	**0.75**	0.56	**0.62**
(E4) to (E4*)	0.47	**0.46**	0.40	**0.45**	0.87	**0.9**
(F3) to (F3*)	0.52	**0.60**	0.50	**0.60**	1.0	**1.0**
(E5) to (E5*)	0.57	**0.78**	0.40	**0.55**	0.71	**0.79**
(E6) to (E6*)	0.59	**0.74**	0.33	**0.56**	1.0	**1.0**
(F4) to (F4*)	0.52	**0.62**	0.33	**0.52**	0.54	**0.71**
(E7) to (E7*)	0.67	**0.79**	0.50	**0.72**	0.71	**0.79**
(F5) to (F5*)	0.54	**0.59**	0.40	**0.51**	0.74	**0.78**
(E8) to (E8*)	0.55	**0.60**	0.40	**0.51**	0.41	**0.48**
(F6) to (F6*)	0.40	**0.45**	0.29	**0.36**	0.43	**0.52**
(E9) to (E9*)	0.45	**0.50**	0.33	**0.47**	0.50	**0.57**
(F7) to (F7*)	0.42	**0.79**	0.60	**0.67**	0.70	**0.74**
(E10) to (E10*)	1.0	**1.0**	1.0	**1.0**	0.83	**0.86**

However, as human judgment provides the most reliable matching, the quality of results is determined within two dimensions. First, the same process elements should be identified as optimal matches by BPS methods. Second, the methods should produce high similarity scores for those matches. Investigating the first requirement, optimal process element combinations produced by human judgment and by all three methods are identical. However, when applying similarity metrics, e.g., for inductive reference modeling, a certain threshold determines if two business process elements are similar or not. Thus, if scores fall below that threshold, corresponding elements are determined as different, which negatively affects the resulting implications. However, regardless of the set threshold, Table 1 shows, that each metric produces better scores with IC than without.

As illustrated in Figure 4, for syntactic similarity, the average similarity score increases from 0.59 without IC to 0.69 if IC is enabled. Analyzing semantic similarity, an increase from 0.46 to 0.59 can be observed. Finally, contextual similarity shows an average similarity score of 0.73, which increases to 0.78 with IC.

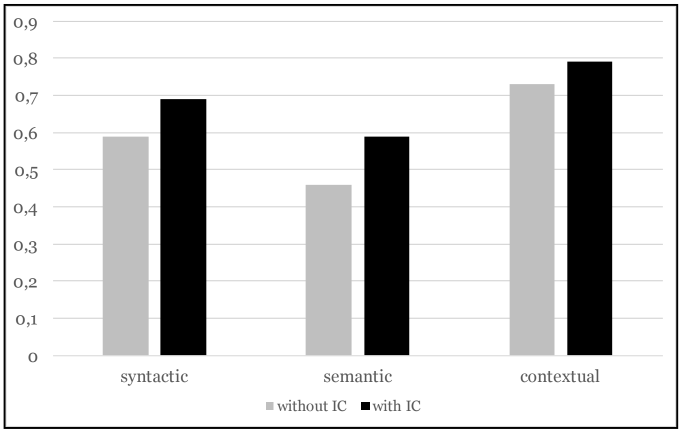

Figure 4: Overview of Average Similarity Scores

Conclusion

The present paper aimed to improve traditional BPS metrics by considering the IC of each word contained by a process element's label. Based on the metrics of syntactic, semantic, and contextual similarity, we provided theoretical foundations as well as an empirical evaluation. Our experimental validation suggests that the quality of results increases noticeably by integrating IC into the underlying calculation schemes. Improvements are mainly because the supposed approach controls for the degree to which a word contributes to the meaning of a label. This is achieved by assigning a specific weighting factor to each word within a label that is equal to its IC which we retrieved from a representational text corpus. Considering our experimental evaluation, metrics are more accurate and reliable with IC than without. Since the similarity of business process elements is the basis for metrics that compare full-scale process models, numerous application domains, such as Process Mining, Business Process Collection Management, and Inductive Reference Modeling can benefit from these findings.

However, approaches of this kind have a variety of well-known limitations. One major drawback results from the statistical noise produced by the used similarity metrics themselves. To generate an equivalence mapping between two business processes, a threshold needs to be defined that determines if two business process elements are equal or not. Since comprehensive empirical validations are not provided within the existing literature, the mapping can be significantly biased by individual perceptions. In line with that, the computation of semantic similarity is based on heuristically determined weighting factors. According to an experimental validation by Dijkman et al. (2011), weighting factors of 1 for equal words and 0.75 for synonyms produce the best results. In addition, the explanatory power of our findings is limited by the usage of GermaNet as a lexical-semantic database. Although GermaNet contains more than 100.000 synsets, its does not sufficiently cover the domain of BPM. Thus, domain-specific vocabulary cannot be identified as synonyms, negatively influencing the metric's applicability. If words cannot be identified within the database, a mapping of process elements is exclusively based on syntactic similarity, not taking any semantic information into account. Future research should focus on the development of a lexical-semantic database that accounts for domain-specific vocabulary. Another issue that was not addressed in this study was the selection of an adequate text corpus. Like GermaNet, the Leipzig Corpora Selection does not provide domain-specific vocabulary and can result in inconsistent IC-scores. Considerably more work will need to be done to determine the influence of a corpus on the overall quality of results. Additionally, limitations result from the experimental evaluation of our work. Although, the introduced hypotheses are experimentally validated, the number of evaluated business process elements is too small to provide generalizable implications. Large randomized controlled trials could provide more definitive evidence. Finally, other techniques of natural language processing could have a significant impact on the results of our approach and must be evaluated in future research.

References

Akkiraju, R. and Ivan, A. "Discovering Business Process Similarities: An Empirical Study with SAP Best Practice Business Processes," in *Service-Oriented Computing: 8th International Conference, ICSOC 2010, San Francisco, CA, USA, December 7-10, 2010. Proceedings*, P.P. Maglio, M. Weske, J. Yang, and M. Fantinato (eds.), Springer, Berlin, 2010, pp. 515-526.

Awad, A. "BPMN-Q: A Language to Query Business Processes," in *Proceedings of the 2nd International Workshop on Enterprise Modelling and Information Systems Architectures*, *Vol. 119*, St. Goar, 2007, pp. 115-128.

Baskerville, R., Pries-Heje, J., and Venable, J. "Soft Design Science Methodology," in *Proceedings of the 4th International Conference on Design Science Research in Information Systems and Technology*, Philadelphia, PA, 2009, pp. 1-11.

Becker, J., Rosemann, M., von Uthmann, C., and Uthmann, C.V. "Guidelines of Business Process Modeling," *Business Process Management* (1806), 2000, pp. 241-262.

Biemann, C., Heyer, G., Quasthoff, U., and Richter, M. "The Leipzig Corpora Collection-monolingual Corpora of Standard Size," *Proceedings of Corpus Linguistic*), 2007.

Cayoglu, U., Dijkman, R., Dumas, M., Fettke, P., García-Bañuelos, L., Hake, P., Klinkmüller, C., Leopold, H., Ludwig, A., Loos, P., Mendling, J., Oberweis, A., Schoknecht, A., Sheetrit, E., Thaler, T., Ullrich, M., Weber, I., and Weidlich, M. "Report: The Process Model Matching Contest 2013," in *Business Process Management Workshops: BPM 2013 International Workshops, Beijing, China, August 26, 2013, Revised Papers*, N. Lohmann, M. Song, and P. Wohed (eds.), Springer International Publishing, Cham, 2014, pp. 442-463.

Dijkman, R., Dumas, M., and García-Bañuelos, L. "Graph Matching Algorithms for Business Process Model Similarity Search," in *Business Process Management: 7th International Conference, BPM 2009, Ulm, Germany, September 8-10, 2009. Proceedings*, U. Dayal, J. Eder, J. Koehler, and H.A. Reijers (eds.), Springer, Berlin,, 2009, pp. 48-63.

Dijkman, R., Dumas, M., van Dongen, B., Krik, R., and Mendling, J. "Similarity of Business Process Models: Metrics and Evaluation," *Information Systems* (36), 2011, pp. 498-516.

Dumas, M., García-Bañuelos, L., and Dijkman, R.M. "Similarity Search of Business Process Models," *IEEE Data Eng. Bull.* (32:3), 2009, pp. 23-28.

Dumas, M., Rosa, M.L., Mendling, J., and Reijers, H.A. *Fundamentals of Business Process Management*, Springer, Berlin, 2013.

Ehrig, M., Koschmider, A., and Oberweis, A. "Measuring Similarity between Semantic Business Process Models," in *Proceedings of the fourth Asia-Pacific Conference on Comceptual Modelling*, Australian Computer Society, Inc., Ballarat, Australia, 2007, pp. 71-80.

Gregor, S. and Hevner, A.R. "Positioning and Presenting Design Science Research for Maximum Impact," *MIS Quarterly* (37:2), 2013, pp. 337-355.

Hamp, B. and Feldweg, H. "GermaNet - a Lexical Semantic Net for German," *Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*), 1997, pp. 9-15.

Hoffmann, W., Kirsch, J., and Scheer, A.-W. "Modellierung mit Ereignisgesteuerten Prozeßketten," in *Veröffentlichungen des Instituts für Wirtschaftsinformatik (IWi), Universität des Saarlandes*, 1993.

Imgrund, F., Janiesch, C., and Rosenkranz, C. "“Simply Modeling” – BPM for Everybody: Recommendations from the Viral Adoption of BPM at 1&1," in *Business Process Management Cases: Digital Innovation and Business Transformation in Practice*, J. vom Brocke and J. Mendling (eds.), Springer, Heidelberg, 2017, pp. 1-20.

Kucera, H. and Francis, W. "A Standard Corpus of Present-day Edited American English, for Use with Digital Computers," Brown University Press, Providence, RI, 1979.

Kunze, M., Weidlich, M., and Weske, M. "Behavioral Similarity – A Proper Metric," in *Business Process Management: 9th International Conference, BPM 2011, Clermont-Ferrand, France, August 30 - September 2, 2011. Proceedings*, S. Rinderle-Ma, F. Toumani, and K. Wolf (eds.), Springer, Berlin, 2011, pp. 166-181.

Levenshtein, V.I. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals," *Soviets physics doklady* (10), 1966.

Li, Y., McLean, D., Bandar, Z.A., Shea, J.D.O., and Crockett, K. "Sentence Similarity Based on Semantic Nets and Corpus Statistics," *IEEE Transactions on Knowledge and Data Engineering* (18:8), 2006, pp. 1138-1150.

Manning, C.D. and Schütze, H. *Foundations of Statistical Natural Language Processing, Vol. 999*, MIT Press, Cambridge, MA, 1999.

Martens, A., Fettke, P., and Loos, P. "A Genetic Algorithm for the Inductive Derivation of Reference Models Using Minimal Graph-Edit Distance Applied to Real-World Business Process Data," *Tagungsband Multikonferenz Wirtschaftsinformatik 2014. Multikonferenz Wirtschaftsinformatik (MKWI-14), February 26-28, Paderborn, Germany*), 2014.

Meadow, C.T., Boyce, B.R., and Kraft, D.H. *Text Information Retrieval Systems*, Academic Press, San Diego, 1992.

Porter, M.F. "An algorithm for Suffix Stripping," *Program: Electronic Library and Information Systems* (14), 1980, pp. 130-137.

Scheer, A.-W. and Nüttgens, M. "ARIS Architecture and Reference Models for Business Process Management," in *Business Process Management: Models, Techniques, and Empirical Studies*, W. van der Aalst, J. Desel, and A. Oberweis (eds.), Springer, Berlin, 2000, pp. 376-389.

Thomas, O. "Das Referenzmodellverständnis in der Wirtschaftsinformatik: Historie, Literaturanalyse und Begriffsexplikation, University of Saarbrücken.

Thomas, O. and Fellmann, M. "Semantische Ereignisgesteuerte Prozessketten.," *Data Warehousing*), 2006, pp. 205-224.

van Der Aalst, W. "Formalization and Verification of Event-Driven Process Chains," *Information and Software Technology* (41), 1999, pp. 639-650.

van Dongen, B., Dijkman, R., and Mendling, J. "Measuring Similarity between Business Process Models," in *Seminal Contributions to Information Systems Engineering: 25 Years of CAiSE*, J. Bubenko, J. Krogstie, O. Pastor, B. Pernici, C. Rolland, and A. Sølvberg (eds.), Springer, Berlin, 2013, pp. 405-419.

Yan, Z., Dijkman, R., and Grefen, P. "Fast Business Process Similarity Search with Feature-Based Similarity Estimation," in *On the Move to Meaningful Internet Systems: OTM 2010: Confederated International Conferences: CoopIS, IS, DOA and ODBASE, Hersonissos, Crete, Greece, October 25-29, 2010, Proceedings, Part I*, R. Meersman, T. Dillon, and P. Herrero (eds.), Springer, Berlin, 2010, pp. 60-77.