Free Text Customer Requests Analysis: Information Extraction Based on Fuzzy String Comparison

Alexander Smirnov, Nikolay Shilov, Kathrin Evers, Dirk Weidig

Free Text Customer Requests Analysis: Information Extraction Based on Fuzzy String Comparison

Alexander Smirnov[1], Nikolay Shilov[1*], Kathrin Evers[2], Dirk Weidig[2]

[1] SPIIRAS, St. Petersburg, Russia
smir@iias.spb.su, nick@iias.spb.su*
[2] Festo SE & Co. KG, Esslingen, Germany
kathrin.evers@festo.com, dirk.weidig@festo.com
*Corresponding author

Abstract. Complex systems (such as automation systems), from here on referred as "products" are usually difficult for customers to specify since there are a lot of parameters to be defined and the customer should perfectly know what is really needed and important for the supplier in order to provide for a proper system. As a result, creating forms and templates for the customer request specification entry helps only for relatively simple tasks and the completely digital request acquisition and processing is still a matter of future work. Currently, the original request specification comes from the customer in various ways (texts, images, diagrams, a phone or a direct talk to the company's sales representative) and the results of the analysis of this specification are often forwarded further to the back-office in a form of free or semi-structured text written in natural language. Since this text is the main source of information about the customer request, it is very important to extract as much information from it as possible. The paper reports the research and development work on semantic text analysis for information extraction from customer requests written in natural language. The core of the work is development of methods for finding a pre-defined list of terms (product parameters that are important for the order specification) in a fuzzy (similarity-based) manner with the help of synonym dictionaries. The results are illustrated on a case study from the automation equipment producer Festo SE & Co KG.

Keywords: Free text analysis · Semantic analysis · Fuzzy string comparison

1 Introduction

Today, product lifecycle management (PLM) automation is devoted a significant attention [1–4]. In previous publications we considered intelligent IT support for the entire PLM cycle [5] and the PLM stage of marketing [6, 7]. One of the possible results of this process is "Digital Customer Journey" [8] standing for complete IT support of the customer at all PLM stages from finding the supplier through product configuration and sales to product usage and disposal (fig. 1).

Fig. 1. Digital customer journey [8].

In this paper we would like to concentrate at the early PLM stage of product configuration. Complex automation systems as a rule are sold on the assemble-to-order basis, i.e. the bill of material and is collected out of built to stock (or sometimes built-to-order) [9] components based on the customer's specification. The problem is that such complex systems are usually difficult for customers to specify since there are a lot of parameters to be defined and the customer should perfectly know what is really needed and important for the supplier in order to provide for a proper system. As a result the complete "digital customer journey" from submitting a request and getting the product to its usage and utilisation is currently possible for relatively simple products and systems. For the considered type of products (automation systems) the "journey" starts from submitting a quotation request by the customer in a format that is from the customer's point of view the most convenient and easiest for understanding, including (but not limited to) texts, images, diagrams, or a phone calls.

The request is addressed to a sales representative or application engineer who specifies the customer requirements that are forwarded further to the back-office in a form of free or semi-structured text written in natural language (fig. 2). For successful customer request fulfilment it is very important to extract as much information from this text as possible. The difficulty of this task is caused by the fact that the texts are filled out manually (sometimes not in native language of the writer), they are very short (e.g., see the text in the tables of fig. 2), mostly contain specific technical terms, and the training set is relatively small (about 800 samples).

The contribution of this paper is the result of the research and development work on semantic text analysis for information extraction from customer requests written in natural language. The work is split into two parts: development of a method for finding a pre-defined list of terms (parameters of the product, parameters of the automation process the product takes part in, characteristics of objects it has to work with, e.g, workpiece characteristics, and other) from the tables in a fuzzy (similarity-based) manner; and development of a method for extracting information from the longer texts of the requests. While the former task is solved, the latter is still the matter of future research. The report does not discuss typical operations for preparing strings for comparison (deleting special characters, converting all characters to lower case and others), since they

Your task

Workpiece	Work pieces made of brass	
max. weight	0.02	kg (assumption)
Dimensions workpiece	different	mm (LxWxH)
Stroke Z-Axis	max. 100	mm
Stroke Y-Axis	max. 300	mm
Stroke X-Axis	appro. 0	mm
Cycle time	1.0	s

Notes
The dimensions of the work pieces are diffenrent. In the process of gripping it has to be ensure that the gripper finger can grip all the pieces.

Your task

Pick up of parts from a moving belt with tracking using a camera system and tripod. The parts are detected with a camera system and gripped by a vacuum gripper or a mechanical gripper.
Due to the different shapes of the parts, a tool change might be necessary to have an accurate gripping solution. The parts are aligned with a rotational drive and placed to a defined position.

In this offer is a vacuum gripper included as an example gripping solution. Further details about the parts to handle will be necessary to select an appropriate gripper.

Work piece		
max. weight	0,05	kg
Dimensions work piece	50x20x2	mm (LxWxH) approximately
Cycle time	10	s (cycle not defined in inquiry)

Fig. 2. Examples of customer request specification in a form of semi-structured text.

are standard. The results are illustrated on a case study from the automation equipment producer Festo SE & Co KG.

The paper is structured as follows. Sec. 2 discusses the state of the art in the area of semantic analysis of short texts. It is followed by the description of the developed method for extraction of information from very short texts (table cells). Sec. 4 applies this method to extraction of information from longer texts. The results are summarized in the conclusion section.

2 State of the Art Review

The authors of [10] deal with a similar task, namely understanding of short texts (search requests). The authors analyse different approaches to text analysis taking into account the specifics of short (several words long) texts. Though, this approach seems to be applicable, the authors use the Probase dictionary [11] for reasoning on text semantics. However, this work deals with various technical terms and we could not find any sufficient dictionaries covering the required terminology. As a result it was decided not to continue with approaches relying on dictionaries for their operation.

In [12] the authors have identified and analysed three major directions of text analysis with regard to analysis of short texts, namely:

- Text segmentation – dividing text into a sequence of meaningful components.
- Part of Speech (POS) tagging – identification of the lexical types (parts of speech) of the words in the text.
- Labelling – definition of the most appropriate concepts meant by a word or a phrase within specific context.

Text segmentation approaches [13] can be split into two groups: approaches treating input text as a set of words and approaches considering phrases (e.g., longest matching methods trying to find the longest known (available in a dictionary) phrase in the text). Working with separate words will not be efficient in the considered domain, since they might have different meaning based on the context. For example, “max payload” would

usually mean the weight of the workpiece only, and "max payload for EXCT" would mean the weight of the workpiece together with the weight of the corresponding gripper or other holding device attached to the 2-dimensional linear gantry EXCT. At the same time, the longest matching methods (e.g., [14]) might be of interest since the longer text in this case can be considered as text template (text construction block) with a concrete meaning. Thus, the principles of these methods are partially used in the proposed approach.

POS tagging algorithms can also be split into two groups: rule-based algorithms (based on linguistic rules) and statistical algorithms (based on accumulated statistics including neural networks) [15]. The authors of [12] note that algorithms of both groups assume that the analysed text is properly structured. However, this is not the case for short texts. For example, the following fragments have the same meaning: "max. workpiece weight", "max. weight of workpiece", "max. weight workpiece", "max. weight/workpiece". As a result, POS tagging was not considered as a possible solution for the task set.

Labelling or concept labelling is aimed at identification of certain terms (concepts) in the text. Usually, such methods are built around statistical models [16, 17] or machine learning techniques [18, 19]. The problem with these methods is in extensive usage of available dictionaries and large training sets. However, as it was mentioned, the current research deals with short texts using specific technical terms. Besides, the available training sets were not large enough for training sophisticated AI models. The work by E. Brill [20] is slightly different. Though also based on statistical data, the result of training is a set of rules. An example of such a rule is: *if a word is tagged as the infinitive "to" and the following word is tagged as an article, than the infinitive "to" is replaced with the preposition "to"*. This can be illustrated as follows:

… *to load* the system – "to load" remains an infinitive;

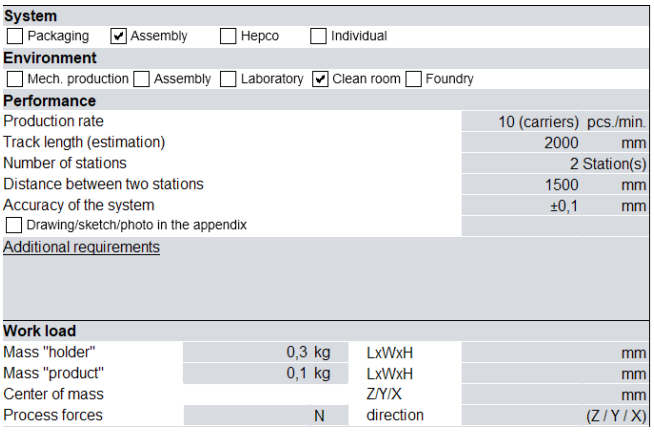
… *to a load* of 1 kg – "to" becomes a proposition and "load" becomes a noun.

Similar rules can be defined for the considered problem in order to better understand the semantic meaning of the words in the short texts analysed.

3 Extraction of Product Parameters from Short Texts

The goal of this task is to set semantic correspondence between texts represented in tables (fig. 2) with known parameters of the product. The reference parameters were extracted from product datasheets (a fragment is shown in fig. 3). Such datasheets are used within the company for product description and can be considered as target models for information extraction. As it can be seen in fig. 2, the text in tables is extremely short and its processing purely by methods described in sec. 2 would not produce any sufficient results. Experimentation with different algorithms showed that that best results could be achieved via combination of fuzzy (similarity-based) string matching with usage of synonyms supported by pre-defined rules.

There are a number of algorithms for evaluating string similarity measures [21]. One of the most popular is the Levenshtein coefficient. It is based on counting the number of character edits (deleting, inserting or replacing one character) that have to be done



System

☐ Packaging ☑ Assembly ☐ Hepco ☐ Individual

Environment

☐ Mech. production ☐ Assembly ☐ Laboratory ☑ Clean room ☐ Foundry

Performance

Production rate	10 (carriers)	pcs./min.
Track length (estimation)	2000	mm
Number of stations	2	Station(s)
Distance between two stations	1500	mm
Accuracy of the system	±0,1	mm

☐ Drawing/sketch/photo in the appendix

Additional requirements

Work load

Mass "holder"	0,3 kg	LxWxH	mm
Mass "product"	0,1 kg	LxWxH	mm
Center of mass		Z/Y/X	mm
Process forces	N	direction	(Z / Y / X)

Fig. 3. Examples of customer request specification in a form of semi-structured text.

to make the compared strings identical. For example, for the strings “max weight of payload” and “max payload weight” (both refer to the weight of the workpiece) the Levenshtein's distance is 16, and for the strings “max weight of payload” and “max payload for EXCT” (the second string refers to the working load of the 2-dimensional linear gantry EXCT) the Levenshtein's distance is also 16. The calculation time for 1000 comparisons of strings “max weight of payload” and “max payload weight” on a computer with Intel Pentium i7 processor is 2.61 ms). The first drawback of this algorithm (difficulty of evaluating / comparing result for strings of different lengths) can be easily overcome, for example, by dividing the result by the length of the larger of the two compared strings (the calculation time increases insignificantly – 3.82 ms under the same conditions). However, the significant drawback demonstrated in the above example is poor results for strings obtained via rearranging words (“max weight of payload” and “max payload weight”), which is a common situation for example for the English language.

As a result of the analysis of existing algorithms for evaluating string similarity, an algorithm was selected based on the calculation of the Sørensen–Dice coefficient for sets [22], which is calculated by the following formula:

$$K = \frac{2|A \cap B|}{|A| + |B|}$$

where A and B are sets consisting of all possible substrings of the strings being compared.

Implementation if this comparison produces the following results:

“max weight of payload” and “max payload weight”: 0.37

“max weight of payload” and “max payload for EXCT”: 0.23

Obviously, this algorithm handles situations related to word rearrangement better than the Levenshtein algorithm. However, the calculation time is significantly higher than that of the Levenshtein algorithm (106.88 ms for 1000 comparisons). To improve the algorithm, experiments were carried out with a change in the length of the substrings

being compared, and as a result, an option was chosen with a comparison of only two-letter substrings as having minimal computational complexity but providing sufficient precision and recall for the test set. The result of the operation of this algorithm on the examples above is:

"max weight of payload" and "max payload weight": 0.94
"max weight of payload" and "max payload for EXCT": 0.53

Such result obviously meets the expectations and the calculation time for 1000 comparisons with the improved algorithm is 11.67 ms.

Tests on the available training set showed good results and indicated that spell check was not required (the algorithm processes misspellings by itself). To improve the matching some synonyms had to be added for the situations when the reference parameter text was significantly different from the text in the training set. For example, for the reference parameter name "working load (front unit + workpiece, effective load for Z-axis)" the following synonyms were added: "total load", "moved mass including gripper", "max weight for", "max payload for", "max weight of loading", etc. With 10 synonyms, the rate of 100% hits was achieved for the complete training set with the threshold value of 0.5.

Analysis of the second column (fig. 2), i.e. extraction of numeric values from texts was done based on pre-defined rules and regular expressions. Since the variety of possible combinations was much smaller, this task did not cause any significant problems, though such cases as "number", "range" (e.g., "3 - 4"), "box" (e.g., "400 x 250 x 200") and "cylinder" (e.g., "d20 x 100") had to be taken care of. Similar solution was implemented for the column with measurement units.

However, though tables are the easiest parts of the requests for processing, still some significant information is contained within the free text accompanying the tables. The next section describes the approach used to extract information from these.

4 Finding Short Text Fragments in Long Text

Extracting semantic from the longer free text describing the customer request is currently an ongoing work. At the beginning of research, a method for discovering keywords corresponding to the mentioned above reference terms from the product datasheets (fig. 3) was developed. The aim of the method was finding the reference terms within the text taking into account possible misspellings, different wordings and usage of synonyms. For this purpose the direct search was chosen assuming similarity measurement between the reference term or its synonym and all possible substrings of the considered text. The similarity measurement is done via the same method as used for short text matching (sec. 3) with the same threshold value of 0.5.

Further in this section we consider an example with the text consisting of 426 symbols. The original text is the following:

All axis could move at the same time up/down. No sidewise movement when Z-axis is in advanced position. Clamping unit required (Pneumatic locking not desired). Sensing required for open clamp. Absolute linear encoder with DRIVE Cliq technology.

Color of steel profile: RAL 7035. Manual greasing of roller bearing cassettes needed. Not allowed fall down the grease of any component. No welded parts (or US certificate if welded parts exist).

The cleaned text is as follows:

all axis could move at the same time up down no sidewise movement when z axis is in advanced position clamping unit required pneumatic locking not desired sensing required for open clamp absolute linear encoder with drive cliq technology color of steel profile ral 7035 manual greasing of roller bearing cassettes needed not allowed fall down the grease of any component no welded parts or us certificate if welded parts exist

The method produced expectable results (all keywords were found successfully, and some additional substrings similar to the keywords were found as well), however, the total calculation time for this text on the Intel Pentium i7 computer took 169.383 s. Obviously, this time is too long for daily usage and a search for optimization of the method has been carried out. One can guess that comparing a short reference term with a long text will not result in high similarity. This has led to an effort to study the dependence of the similarity level on the length of the compared strings. The relative length of the compared substring was measured as

$$rl = \frac{length(substring)}{length(ref\_term)}$$

where *ref_term* is reference term being compared, and *substring* is the substring being compared.

The result of experiments with reference terms having different lengths is shown in fig. 4. The magnified fragment for the relative length between 0 and 10 is shown in fig. 5. It can be seen that for substrings with relative length more than 3.5 the similarity does not exceed the threshold. This means that limiting the length of the matched substring as 3.5 times the length of the reference term will not cause any loss of possibly matching substrings if the matching threshold is 0.5. On the other hand the calculation time with this limit is 1.469 and the produced result is the same.

As it can be seen, at the moment proposed method is oriented only to one task (finding reference terms in the text) and does not take into account semantics. Planned future work will include using of different semantic-based free text analysis methods including word / phrase labelling.

5 Conclusion and Future Work

The paper considers the problem of customer support at the product configuration PLM stage for assemble-to-order inventory method. Efforts towards achieving a complete support of the digital customer journey (when the customer is supported by IT services at all stages of PLM) have set new problems to be solved. In this paper we consider the problem of semantic text analysis for information extraction from customer requests written in natural language. The work is split into two parts: development of a method for finding matching very short texts (1-5 words) to a pre-defined list of terms (product parameters that are important for the order specification) in a fuzzy (similarity-based)

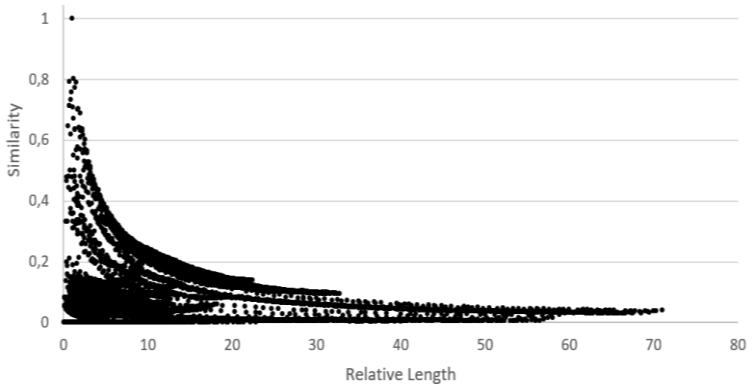


Fig. 4. Dependence of the similarity level on the relative length of the compared text fragment.

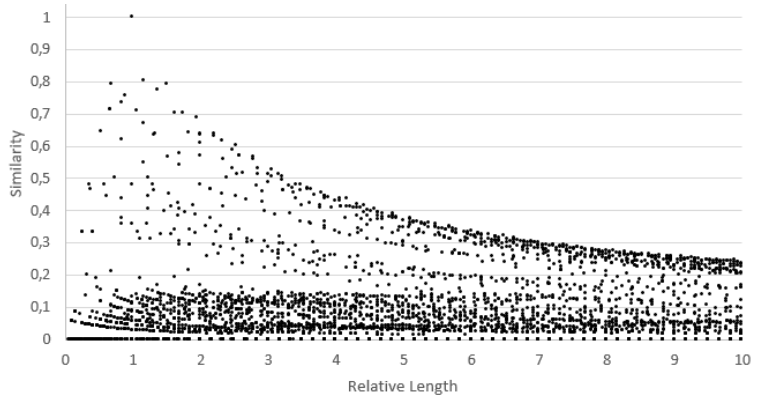


Fig. 5. Dependence of the similarity level on the relative length of the compared text fragment: magnified fragment for relative length between 0 and 10.

manner; and development of a method for extracting information from the longer texts describing the customer requests.

The first task has been solved through combination of fuzzy (similarity-based) string matching based on the Sørensen–Dice coefficient supported by usage of pre-defined rules and synonyms. The method gave 100% precise matching results on the training set. Achieving such result was possible due to exhaustive analysis of the customer requests and collecting the list of synonyms. Obviously for new requests the matching result will be lower. For this purpose a tool for company experts to evaluate and extend this list of synonyms will be developed.

Tackling the second task has been started with combining the above method with direct search and further optimisation to reduce the calculation time. Planned future work includes using of different semantic-based free text analysis methods including word / phrase labelling.

Acknowledgements. The paper is due to collaboration between SPIIRAS and Festo SE & Co KG (the case study), grant # 18-07-01203 of the Russian Foundation for Basic Research (the framework and methods), and State Research no. 0073-2019-0005 (analysis of customer needs at PLM stages).

References

1. Ferreira F, Faria J, Azevedo A, Marques AL (2017) Product lifecycle management in knowledge intensive collaborative environments: An application to automotive industry. International Journal of Information Management 37:1474–1487. https://doi.org/10.1016/j.ijinfomgt.2016.05.006
2. Holligan C, Hargaden V, Papakostas N (2017) Product lifecycle management and digital manufacturing technologies in the era of cloud computing. In: 2017 International Conference on Engineering, Technology and Innovation (ICE/ITMC). IEEE, pp 909–918
3. Liu Y, Zhang Y, Ren S, et al (2020) How can smart technologies contribute to sustainable product lifecycle management? Journal of Cleaner Production. https://doi.org/10.1016/j.jclepro.2019.119423
4. Abramovici M, Göbel JC, Savarino P, Gebus P (2017) Towards Smart Product Lifecycle Management with an Integrated Reconfiguration Management. In: Product Lifecycle Management and the Industry of the Future. IFIP Advances in Information and Communication Technology. Springer, pp 489–498
5. Smirnov A, Shilov N (2018) Multi-aspect Ontology for Semantic Interoperability in PLM: Analysis of Possible Notations. In: IFIP International Conference on Product Lifecycle Management. PLM 2018: Product Lifecycle Management to Support Industry 4.0. Springer, pp 314–323
6. Smirnov A, Shilov N, Oroszi A, et al (2017) Changing Information Management in Product-Service System PLM: Customer-Oriented Strategy. In: Ríos J, Bernard A, Bouras A, Foufou S (eds) IFIP Advances in Information and Communication Technology. PLM 2017: Product Lifecycle Management and the Industry of the Future. Springer, pp 701–709
7. Krebs T, Oroszi A, Sinko M, et al (2018) Changing information management for product-service system engineering: customer-oriented strategies and lessons learned. International Journal of Product Lifecycle Management 11:1–18. https://doi.org/10.1504/IJPLM.2018.10012695
8. Oroszi A (2019) Digitalization at Festo - Our way in digital transformation. Keynote speech. In: 9th IFAC Conference MIM 2019
9. Smirnov A, Kashevnik A, Shilov N, et al (2015) Changing Business Information Systems for Innovative Configuration Processes. In: Matulevičius R, Maggi FM, Küngas P (eds) Joint Proceedings of the BIR 2015 Workshops and Doctoral Consortium co-located with 14th International Conference on Perspectives in Business Informatics Research (BIR 2015). CEUR, pp 62–73
10. Hua W, Wang Z, Wang H, et al (2017) Understand Short Texts by Harvesting and Analyzing Semantic Knowledge. IEEE Transactions on Knowledge and

Data Engineering 29:499–512. https://doi.org/10.1109/TKDE.2016.2571687
11. Wu W, Li H, Wang H, Zhu KQ (2012) Probase: a probabilistic taxonomy for text understanding. In: Proceedings of the 2012 international conference on Management of Data - SIGMOD '12. ACM Press, New York, New York, USA, p 481
12. Hua W, Wang Z, Wang H, et al (2015) Short text understanding through lexical-semantic analysis. In: 2015 IEEE 31st International Conference on Data Engineering. IEEE, pp 495–506
13. Pak I, Teh PL (2018) Text Segmentation Techniques: A Critical Review. In: Innovative Computing, Optimization and Its Applications. Studies in Computational Intelligence. Springer, pp 167–181
14. McDonald R, Crammer K, Pereira F (2005) Flexible text segmentation with structured multilabel classification. In: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05. Association for Computational Linguistics, Morristown, NJ, USA, pp 987–994
15. Kumawat D, Jain V (2015) POS Tagging Approaches: A Comparison. International Journal of Computer Applications 118:32–38. https://doi.org/10.5120/20752-3148
16. Zhou G, Su J (2001) Named entity recognition using an HMM-based chunk tagger. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02. Association for Computational Linguistics, Morristown, NJ, USA, p 473
17. McCallum A, Li W (2003) Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 -. Association for Computational Linguistics, Morristown, NJ, USA, pp 188–191
18. Song Y, Wang H, Wang Z, et al (2011) Short Text Conceptualization Using a Probabilistic Knowledgebase. In: Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence Short. AAAI Press, pp 2330–2336
19. Kim D, Wang HW, Oh A (2013) Context-dependent conceptualization. In: Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. AAAI Press
20. Brill E (1992) A simple rule-based part of speech tagger. In: Proceedings of the workshop on Speech and Natural Language - HLT '91. Association for Computational Linguistics, Morristown, NJ, USA, pp 112–116
21. Kysela J (2018) A Comparison of Text String Similarity Algorithms for POI Name Harmonisation. In: Articulated Motion and Deformable Objects. Lecture Notes in Computer Science. Springer, pp 121–130
22. Udagawa Y (2013) Source Code Retrieval Using Sequence Based Similarity. International Journal of Data Mining & Knowledge Management Process 3:57–74. https://doi.org/10.5121/ijdkp.2013.3404