UNIVERSITÀ DI PISA

DIPARTIMENTO DI INGEGNERIA DELL'ENERGIA DEI SISTEMI
DEL TERRITORIO E DELLE COSTRUZIONI

RELAZIONE PER IL CONSEGUIMENTO DELLA
LAUREA MAGISTRALE IN INGEGNERIA GESTIONALE

Unveiling Finite State Machines (FSMs) from Patents with Natural Language Processing (NLP): The Container Farming case study

SINTESI

RELATORI

Prof. Ing. Gualtiero Fantoni, PhD
Dipartimento di Ingegneria Civile e Industriale

Ing. Vito Giordano, PhD
Dipartimento di Ingegneria dell'Energia, dei Sistemi,
del Territorio e delle Costruzioni

IL CANDIDATO

Marco Consoloni
m.consoloni1@studenti.unipi.it

Sessione di Laurea Magistrale del 26/04/2023

Unveiling Finite State Machines (FSMs) from Patents with Natural Language Processing (NLP): The Container Farming case study

Marco Consoloni

Sommario

L'analisi dei brevetti è il processo di analisi dei documenti brevettuali e delle informazioni derivanti dal ciclo di vita degli stessi. Il presente lavoro di tesi si propone di analizzare come il funzionamento dei dispositivi brevettati è espresso attraverso il linguaggio naturale all'interno dei documenti brevettuali. Sono stati utilizzati i modelli concettuali noti in letteratura delle Macchine a Stati Finiti e dei Diagrammi di Attività in congiunzione con alcune tecniche di *Natural Language Processing* (NLP) per distinguere gli aspetti strutturali/statici delle invenzioni da quelli comportamentali/dinamici. I risultati sperimentali mostrano che il sistema di NLP proposto in questo lavoro è in grado di distinguere tra gli aspetti sopra menzionati e, inoltre, fornisce evidenza che i brevetti presentano diverse strutture dal punto di vista statico-dinamico. Il presente lavoro contribuisce a fare luce sul legame tra i linguaggi di modellazione e il linguaggio naturale. Ulteriori indagini in questa area possono rendere possibile l'utilizzo di strumenti di NLP per estrarre modelli comportamentali dei dispositivi brevettati (e.g., FSM o diagrammi di attività) dalle loro descrizioni testuali (e.g., le descrizioni dei brevetti) e fornire, quindi, un riferimento per la digitalizzazione dei prodotti/servizi nel contesto dell'Industria 4.0.

Abstract

Patent analysis is the process of analysing patent documents and other information from patents lifecycle. This thesis work aims to analyse how the functioning of patented devices is expressed in patent descriptions. I use the conceptual models of Finite State Machines (FSMs) and Unified Modeling Language (UML) Activity diagrams in conjunction with Natural Language Processing (NLP) techniques to distinguish structural/static aspects of inventions from behavioural/dynamic ones. The experimental results show that my NLP system is able to distinguish between the aforementioned aspects and, moreover, it provides evidence that patents exhibit different structures from a static-dynamic point of view. The present work contributes to shedding light on the link between modelling language and natural language. Further investigations in this area can make it possible to leverage NLP tools for extracting behavioural models of patented devices (e.g., FSMs or Activity diagrams) from their textual descriptions. This, in turn, may offer the groundwork for product/service digitalization within the industry 4.0 landscape.

1. Introduction

To protect inventions, patent documents describe them as black-box system, deliberately concealing the internal dynamics of patenting devices as much as possible. In this context, modelling the functioning of patented devices by manually extracting information contained in patent documents is a complex and challenging task. In this work, I attempt to extract the behaviour of patented devices (i.e., the functioning of patented devices over time) by analysing their patent descriptions with software-based tools. Based on the formal models of Finite State Machines (FSMs) and Unified Modeling Language (UML) Activity Diagrams, I develop a keyword-based Natural Language Processing (NLP) system to distinguish those sentences which provide dynamic aspects of inventions from those ones describing structural features of patented devices (such as size, materials, geometrical features of components) within patent descriptions. Three Research Questions (RQs) are addressed in this work: **RQ1:** *Do patent descriptions specify both structural and behavioural aspects of patented devices?* **RQ2:** *How static and dynamic-related contents are organized within patent descriptions? Is there any quantitative and qualitative evidence of recurrent sections/patterns in patent documents?* **RQ3:** *Which are the limits of the keyword-based NLP system at distinguishing structural and behavioural aspects of inventions in patents?* To verify the effectiveness and performance of the proposed approach, a case study is conducted on patents in the field of Container Farming.

2. Related Work

The three pillars that are relevant both theoretically and methodologically for the purposes of this work are: (1) **Systems Modelling** is a process in which technical methods are used to create an abstract representation of a system (i.e., model) aimed at providing valuable understanding of the subject being modelled (Pidd, 2004). (2) A **FSM** is a conceptual model typically adopted to effectively represent the behaviour of systems using the concept of states, event and transitions. (3) **NLP techniques** are largely used to extract information from patent documents using software-based tools. To date, several approaches have been used to extract entities such as components (Cascini et al., 2007), technical problems (Liang and Tan., 2007), functions (Fantoni et al., 2013), materials, physical flows and states (Chen et al. 2020). The current state-of-the-art literature focus on extracting specific entities from text to identify the constitutive elements of patented inventions. However, previous studies do not look at patented devices as systems and they end up extracting a set of disconnected entities concerning inventions. On the contrary, I apply the concepts of **modelling perspectives** (Krogstie, 2012; Opdahl and Sindre,

1995) to patented device and I attempt to mine the behaviour of patented device from patent descriptions by exploiting the **modelling constructs** of Harel FSMs (Harel, 1987) and UML activity diagram (Jacobson and Booch, 2021).

3. Conceptual Framework

Based on the related literature, I develop a conceptual framework which aims to provide a conceptual baseline useful to organize the presentation of the methodology proposed in this study. Patented devices can be ontologically regarded as systems, and, consequently, patent descriptions can be seen as descriptions of systems in written form (i.e., in natural language). I propose a binary approach to analyse inventions in which patented devices can be analysed from (1) a "**static viewpoint**" focusing on the architecture of patented devices (e.g., geometric features, component materials and attributes, couplings between elements) and other descriptive features, and (2) a **"dynamic viewpoint"** focusing on the dynamics of patented devices expressed by functions, activities, events and conditions on activity execution. Patent descriptions contain both viewpoints. Linguistically, the former can be expressed by sentences such as *"[0046] The overall size of the container is approximately 8′×8′×20′ or 40′ in length."* (US10219447B1), whereas the latter by sentences such as "*after the liquid supply is completed, close the first pump, and then close the valve on the liquid supply pipe.*" (CN106900391A). Patent descriptions offer a linguistic translation of conceptual models (such as FSMs and UML activity diagrams) used to design inventions. Conceptual models concisely specify the functioning of inventions. Conversely, natural language is not able to clearly distinguish between structural and behavioural aspects of systems. In fact, in patent descriptions it happens that a set of disconnected statements about the device functioning is generated, but the overall functioning of the device is not clearly specified (i.e., the boundaries among the static and dynamic viewpoints are often blurred).

4. Methodology

To respond to the RQs, I developed a keyword-based NLP system for recognizing keywords and multi-words related to the static and dynamic natures of patented devices within patent descriptions. As shown in Figure 1, the methodology is composed of three phases: (1) **Data collection**, where a set of patents are retrieved for the technology of concern. (2) **Keywords extraction with NLP**, where a field-specific dictionary is used in a keyword-based NLP pipeline to extract keywords/expressions from patent descriptions at sentence level, and (3) **post-processing**, where the output of the NLP pipeline is used to calculate a score (Dynamic Score) for each sentence and to develop a representation of the structure of patents from a static-dynamic point of view (Patent Topography). The description of these phases is reported more in detail in section 5 (within the case study), in section 4.1 and 4.2, respectively.

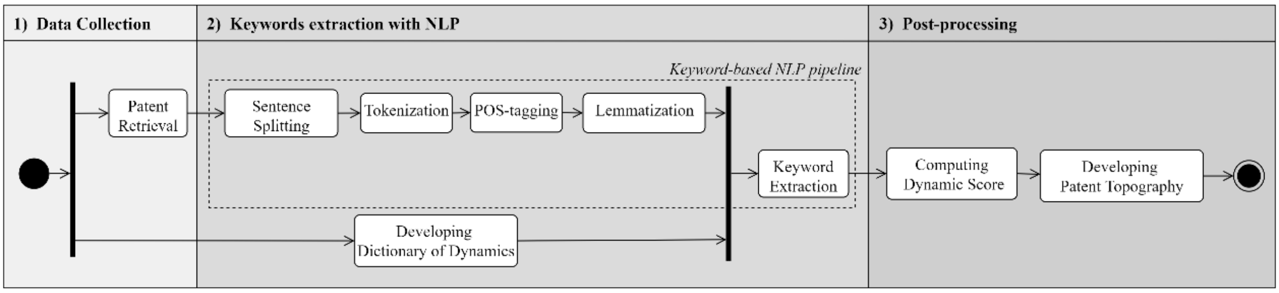

Figure 1. Methodology Workflow.

4.1. Keywords extraction with NLP

I develop a lexicon (Dictionary of Dynamics) with a **hierarchical structure** composed of 496 **keywords and multi-words** related to the dynamic and static natures of patented devices (see Table 1). The lexicon is structured on the highest level into 2 **macro-categories**. The dynamic macro-category groups keywords and multi-words which linguistically express the behaviour of inventions (dynamic keywords) and it is composed of 6 categories. These categories map the UML activity diagram constructs and the keywords which belong to these categories aims at "translating" the modelling constructs of UML activity diagrams into words. Conversely, the static macro-category groups keywords and multi-words which can be used to describe structural aspects of patented devices (static keywords) and it is composed of 4 categories. These categories have been generated based on the structural and attributive viewpoints proposed by Jang and Yoon (2021). Moreover, I added the "Patent jargon" category in the Static part to map the juridical keywords and multi-words used in patent documents.

Table 1. The Architecture of the Dictionary.

Macro-Category	Category	Examples of Keywords and Multi-words	#
Dynamic	Flow of Control	consequently, earlier, in turn, as a result, once.	73
	Loop	cyclically, continuously, over and over, recurrently, iteratively.	14
	Decision Node	in the case, depending upon, whether, even if, provided that.	39
	Parallel Execution	simultaneously, concurrently, in the time of, in sync, all at once.	20
	Functional Verbs	separate, actuate, move, convert, turn.	204
	General Dynamic *	if, without, when, during, since.	22
Static	Structural	include, comprising, part of, consisting, assembling.	13
	Figure Descriptor	drawings, depict, represent, see, shown.	16
	Patent Jargon	claiming, file, but not exclusively, infringe, without wishing to limit.	61
	General Static *	detail, describe, appreciate, discuss, regard.	34
Total			**496**

* Dedicated category for polysemantic keywords and multi-words.

To create the dictionary, I recursively generated the synonyms for the keywords and multi-words with a generative Artificial Intelligence (AI) developed by the research company OpenAI[1]. The dictionary is used in an NLP pipeline which consists of the following steps. First, patent descriptions are split into sentences (**Sentence Splitting**). Then, each sentence is segmented in orthographic units called tokens (**Tokenization**). Each token is morphologically analysed and marked up as corresponding to a set of part of speech tag (e.g., noun, verb, article) (**POS-tagging**). Each POS-tagged token is associated to its lemma (**Lemmatization**). Eventually, the keywords and multi-words contained in the dictionary are searched within the tokens of each sentence and they are returned along with their, Category and Macro-Category (**Keyword Extraction**). I carried out the NLP pipeline using **SpaCy**, a free open-source library for NLP in Python[2].

4.2. Post-processing

To address the RQs, I calculate a score for each sentence, called **Dynamic Score,** based on the occurrences of dynamic and static keywords/multi-words extracted with the NLP pipeline. The formula for calculating the Dynamic Score at sentence level is:

$$Dynamic\ Score = \frac{(N.Dynamic\ Keywords - N.Static\ Keywords\ )}{Sentence\ Lenght}$$

The higher the Dynamic Score, the higher the difference between the number of dynamic and static keywords is. Thus, when the Dynamic Score get a high value, sentences are more likely to describe the behaviour, or functioning, of patented devices and vice versa. The term *Sentence length* normalize the Dynamic Score and allows the comparison between sentences of different lengths. The metric weights each keywords/multiword equally. Thus, when a sentence contains the same number of static and dynamic keywords/multi-words, the Dynamic Score get the value of 0. When no keywords/multi-words of the dictionary are matched within a sentence by the NLP pipeline, the Dynamic Score is set to the value 0. Based on the values of the Dynamic Score, sentences are tagged as shown in Table 2.

Table 2. Sentence Tags based on Dynamic Score on sentence.

Dynamic Score value	Any match found?	Sentence Tag
Positive	Yes	Dynamic
Negative	Yes	Static
Zero	Yes	Intersection
Zero	No	Not tagged

[1] The web page of the AI bot is available at: https://platform.openai.com/overview

[2] The web page of the spaCy library is available at https://spacy.io/

The "**intersection**" tag means that the dynamic content of that sentence may be ambiguous and blurred. The "**Not tagged**" tag provides the information that a sentence has not been map by the dictionary.

To address RQ2, I analyse how sentence tags of Table 2 form sequences of different length within patent descriptions. In this study, a **sequence** is a batch of consecutive sentences which share the same sentence tag[3]. If a patent is made of long dynamic sequences, then it concentrates the description of the functioning of the invention in homogeneous sections of text. Conversely, smaller sequences can be seen as an indicator of fragmentation, meaning that, the natural language switches between static and dynamic viewpoint without much continuity.

5. Case Study: Container Farming Technology

To clarify the application of the methodology and to demonstrate its validity I apply the methodology in a case study on the technological domain of **Container Farming (CF)**. The term CF refers to a farming technique usually composed of a vertical farming system housed by a shipping container [4]. The CF domain has been assessed as a suitable case study for the following reasons: (1) CF patents are expected to describe both the hardware infrastructure of the patented devices (static aspects) and the process of growing crops (dynamic aspects), (2) as an emergent technology with no dominant designs, it requires detail descriptions of patented inventions and (3) my personal knowledge of the domain may ease in the interpretation and validation of the experimental results.

5.1. Data collection

I retrieved a set of patents on CF as input for the methodology proposed in this study. The patents were retrieved from the patent database of the European Patent Office (EPO)[5]. The query used to retrieve the set of patents was carried out with Espacenet[6]. The "backbone" of the query architecture consists of two concepts, that are (1) case-shaped structures (Container)

[3] For instance, considering the set of consecutive sentences tagged as "Dynamic-Static-Static-Static-Dynamic-Intersection-Intersection-Not_tagged", five sequences of different length are identified: Dynamic (with length 1), Static (with length 3), Dynamic (with length 1), Dynamic (with length 2) and Not_tagged (with length 1).

[4] To have a better understanding of the container farming technology I recommend reading the description of **Greenery-s**, a popular container farming technology produce by Freight Farms, available at https://www.freightfarms.com/greenery-s. Here also follows an explainer video of the product: https://www.youtube.com/watch?v=No-jB8217_Y

[5] The official website of European Patent Office (EPO) is available at: https://www.epo.org/

[6] Espacent is an internet-based patent document search service of the EPO available at: https://worldwide.espacenet.com/patent/

and cultivation process (Farming). The query was performed on 30/11/2022 and it retrieved **99 patent descriptions** along with their metadata.

5.2. Experimental results and discussions

I analysed the experimental results of the case study to answer the RQs. The following analyses have been carried out in R programming language using the **IDE RStudio.**

5.2.1. Analysis of Dynamic Score distribution

To address RQ1, I analysed the distribution of the Dynamic Score on 30,282 sentences from patent descriptions (after filtering out "Not tagged" sentences), as shown in Figure 2.

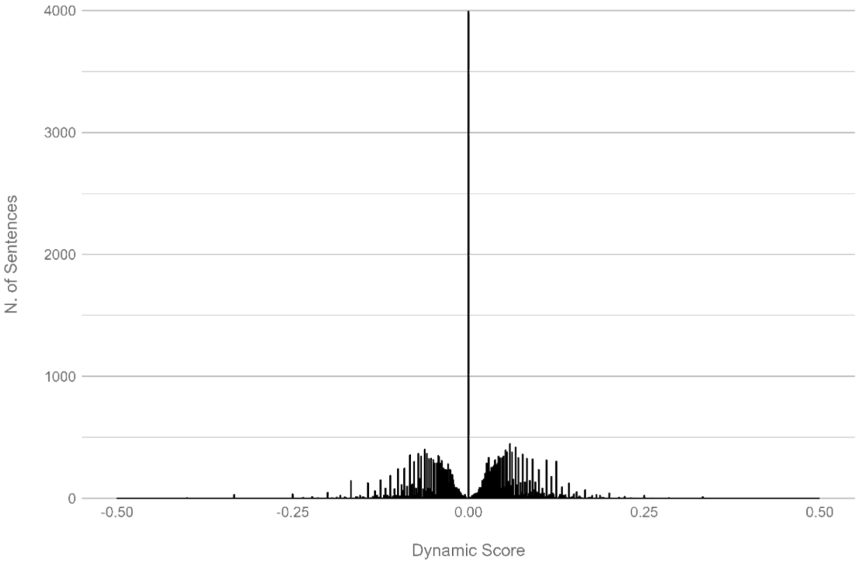

Figure 2. Distribution of the Dynamic Score for sentences from patent descriptions.

The distribution of the Dynamic Score clearly shows a spike over the value 0, meaning that a large number of sentences are tagged as "**Intersection**" (see Table 2). Moreover, the histogram has two peaks of equal height, indicating that the data has two separate regions of higher absolute frequency. The quantiles of the distribution (1st Qu., -0.042; Median, 0.000; 3rd Qu., 0.005) indicates that these two regions are almost symmetric and equally spaced from the centre of the histogram. The valley between the two peaks, as well as the tails of the two regions, offers quantitative evidence supporting the idea that patent descriptions contain sentences with two distinct degrees of static and dynamic-related contents. To give further evidence of this, I report some examples of sentences along with their Dynamic Score in Table3. In the column "Sentence", the dynamic keywords/multi-words matched by the dictionary are in bold font, while static keywords are underlined.

Table 3. Sample of sentences from patent descriptions along with their Dynamic Score and Sentence Tag.

Patent	Sentence	Dynamic Score on sentence	Sentence Tag
GB1083550A	**after closing** and **sealing** the **filled** containers are **then ready for** shipment.	0.500	Dynamic
KR20210078464A	it can be **checked periodically**.	0.400	Dynamic
US2015208592A1	the modules are **supported** and **rotated** in a horizontal position as they are **moved**.	0.286	Dynamic
WO2012072273A1	**then** 45 minutes later row 2 is **turned on**. 45 minutes **later** row 3 may be **turned on.**	0.278	Dynamic
AU2016258913A1	signs of nematode damage include stunting and yellowing of leaves and wilting of the plants **during** hot periods.	0	Intersection
CA2947752A1	each side tab includes a slot **aligned** which receives a colour coded, tamper-proof slide connector.	0	Intersection
KR20180074665A	the second delivery device 235 includes a roller (not shown) that is driven to **move** the container 200 from the second delivery device 235 to the crop storage area 240.	0	Intersection
US11160223B2	the ambient air may comprise humid air.	- 0.143	Static
KR20180074665A	the building 100 includes a frame 110 mounted on a foundation 120.	- 0.167	Static
US2014283452A1	embodiments include light assemblies comprising both fluorescent lamps and leds.	- 0.200	Static
WO2022016291A1	the drawings are not intended to limit the scope of the teachings described herein.	- 0.357	Static

Table 3 provides evidence that the NLP systems correctly capture the static-dynamic “sentiment” of sentences. To further addressing RQ1, I analysed the distribution of Dynamic Score for sentences belonging to specific sections of patent descriptions. I analysed the distribution of the Dynamic Score on sentences belonging to a) the “Background of the Invention” section and b) the “Description of the Drawings” one (see Figure 3).

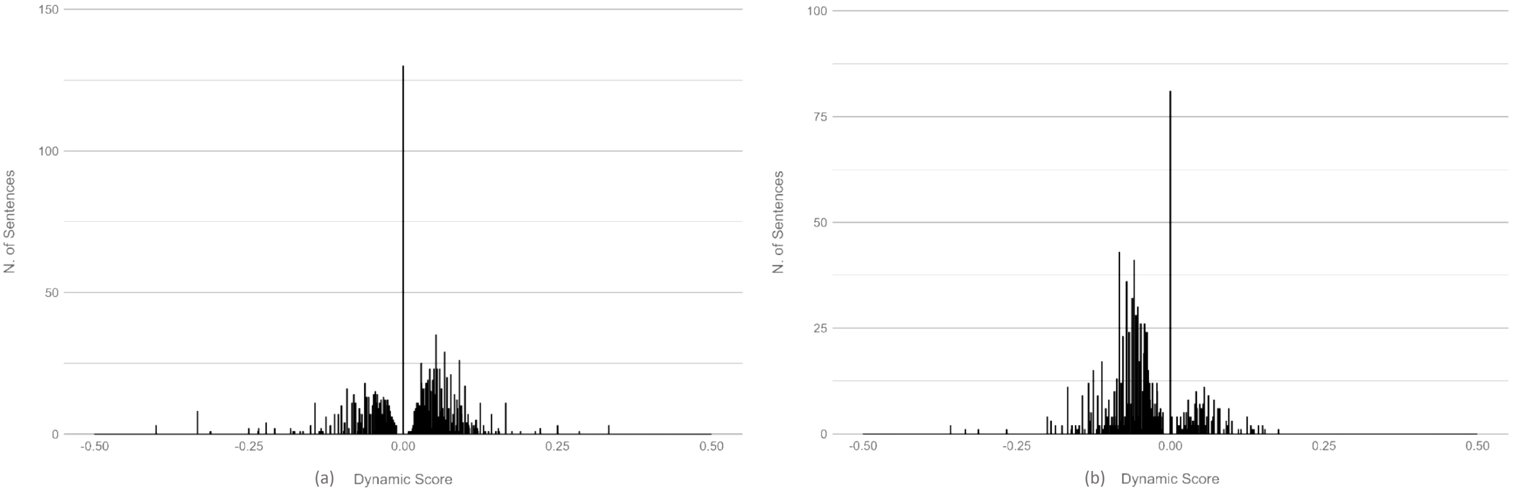

Figure 3. Distributions of the Dynamic Score on sentences belonging to a) the “Background of the Invention” section and b) the “Description of the Drawings” one.

The distributions of the Dynamic Score calculated on these sections shows a comparable pattern with the distribution obtained on the whole patent descriptions. Hence, the experimental results provide quantitative evidence that patent descriptions, as well as its sections, specifies both structural and behavioural aspects of patented devices. Thus, the RQ1 finds a positive answer.

5.2.2. Patent Topography analysis

To address RQ2, I analysed the distribution of **sequence length** for each sentence tags (see Table 2) within patent descriptions. Table 4 reports the statistics for each type of sequence.

Table 4. Summary of statistics about sequence length for each type of sequence

Sequence Type	Min.	1st Qu.	Median	3rd Qu.	Max	Mean	Std	N	%
Dynamic	1	1	1	2	76	1.988	2.409	7,137	30.0
Static	1	1	1	2	60	1.739	1.834	6,959	29.3
Intersection	1	1	1	1	11	1.183	0.672	3,374	14.2
Not tagged	1	1	1	2	82	1.530	1.683	6,305	26.5
Total								**23,775**	**100.0**

The four distributions under consideration exhibit very similar statistical measures. The four sequence types account for equivalent proportions of the total (column "%") and they have highly comparable quantiles, which implies that the distributions share similar patterns of variability and central tendency. Moreover, the mean length of all sequence type is between 1 and 2, indicating that sequences are composed of 1 or 2 sentences on average. This provides quantitative evidence that patent descriptions are typically composed of short sequences supporting the idea that the content related to static and dynamic aspects of patented devices is fragmented by natural language and not confined to distinct and homogeneous sections of text. For these reasons, RQ2 yields a negative answer, as evidence suggests that structural and behavioural aspects of patented devices are distributed throughout patent descriptions.

To gain a more comprehensive understanding on how static and dynamic-related contents are organized within patent descriptions, I calculated the Average Dynamic Score (ADS) for each patent of the dataset. The ADS for a patent is calculated as the sum of the Dynamic Score of its sentences divided by the total number of sentences of the patent. Hence, I selected those patents with the highest (lowest) ADS in order to investigate patents with more dynamic-related (static-related) content. Then, I generated a visualization of the structure of patents from a static-dynamic point of view (**Patent Topography**). To give an example, as shown in Figure 4, I plot for each sentence (x-axis) of two different patent descriptions their Dynamic Score (y-axis). When the Dynamic Score of a sentence is 0, which means that the sentence tag is either "Intersection" or "Not tagged" (see Table 2), then the plot does not show any bars. Figure 4 provides evidence that the two patents have different structures, and that structural and behavioural aspects of inventions are basically expressed throughout patent descriptions according to the writing style of patent authors. In fact, after manually reading the two patents, it is possible to observe that the first one, which contains much more static sentences than dynamic ones, describes in detail the original features of the embodiments of the cultivation system with many references to drawings, whereas the latter focuses more on the dynamics of the cultivation process rather than on the hardware infrastructure of the patenting device.

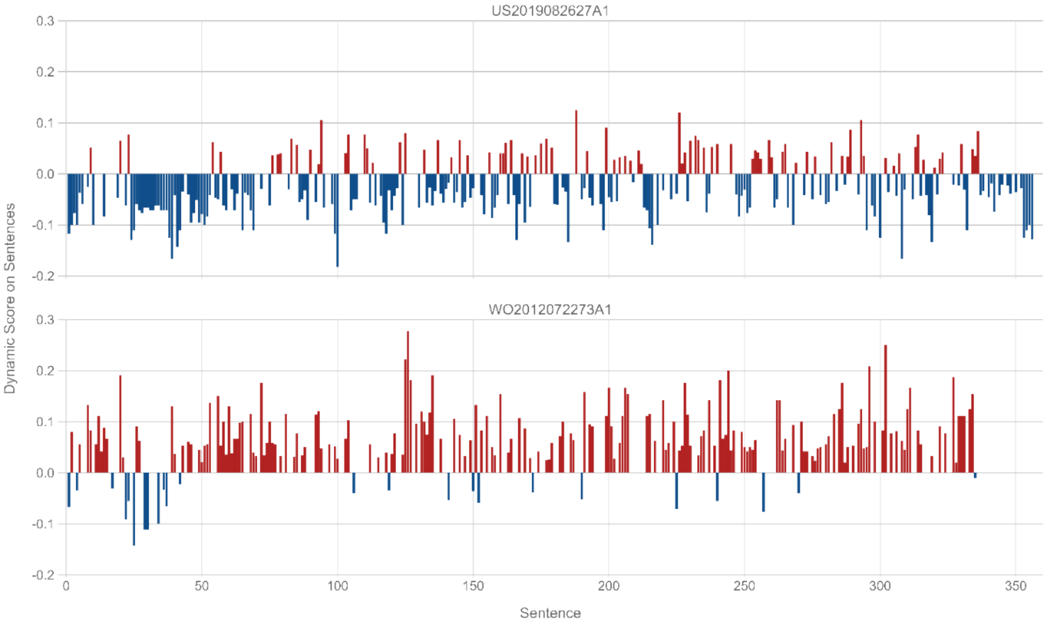

Figure 4. Patent Topography of two patents: 1) US2019082627A1 and 2) WO2012072273A1.

5.2.3. Performance evaluation of the Dictionary

To address RQ3, I evaluate the performance of the dictionary used in the NLP pipeline. To the purpose of this study, I assessed two key metrics: 1) the **dictionary coverage** which measures the percentage of sentences that finds at least a match in the set of keywords and multi-words of the dictionary and 2) the **dictionary overlap** which measures the percentage of sentences that contain both static and dynamic keywords. The first metric measures the effectiveness of the dictionary for the NLP task. The second metric measures the "quality" of the dictionary's keywords. In fact, a low dictionary overlap means that static and dynamic keywords are mutually exclusive (i.e., do not co-occur within sentences), and so, they clearly distinguish between structural and dynamic-related aspects of patented inventions.

Table 5. Performance of the dictionary.

Any Static Keywords matched?	Any Dynamic Keywords matched?	N. of Sentences	Percentage %
NO	NO	9,644	24.15*
NO	YES	10,961	27.46
YES	NO	9,207	23.06
YES	YES	10,114	25.33
Total		**39,926**	**100.00**

* Percentage of "Not Tagged" sentences

The coverage and the overlap of the dictionary are shown in Table 5. The coverage of the dictionary is 100.00 % - 24.15 % = 75.85 %. The overlap of the dictionary is 25.33 %. I analysed (1) which kind of sentences are not matched by any of the dictionary's keywords and (2) which kind of sentences contain both static and dynamic keywords. This helps me in identifying the weaknesses of the NLP approach, as well as its possible improvements. In the first case, I spotted the following main issues: **Domain-specific verbs** (e.g., mill, pipped out, inoculate) and **generic verbs** (to use, to be, to provide) which are not present in the dictionary lower the dictionary coverage. In the second case, I spotted the following main issues: **Dependent clauses** of sentences which allow to combine static and dynamic aspects of patented inventions within sentences and **Polysemy of words** which causes dynamic keywords to be used in static way and vice versa. In both cases, I spotted the following main issues: **Sentence splitting errors** which occurring in the NLP pipeline and text is spilt into short sentences with no actual meaning, or it is mis-splitted into (too long chunks of text). **POS-tagging errors** which occurs when POS-tagging module fails in tagging functional verbs. For those readers more familiar with NLP techniques, **possible improvements** of these issues are reported in the Appendix.

6. Conclusions

The main findings of this work provide quantitative evidence that: (1) patent descriptions specify both structural and behavioural aspects of inventions and, when these aspects are formulated in natural language, the boundaries among them are often blurred. (2) the organization of structural and behavioural aspects of inventions throughout patent descriptions is heavily influenced by the writing style of patent authors and patents exhibit different structures (topographies) from a static-dynamic point of view. (3) NLP has proven to be suitable in distinguishing between these two aspects. The strongest limitation lies in the evaluation of the precision and recall of the NLP system in tagging sentences (supervised classification approach). This depends on the huge efforts needed for developing a set of labelled sentences and the absence of unique interpretation of structural and dynamic aspects of a system. Companies that leverage technical information contained in patent documents to enhance their business strategy and R&D activities may benefit from this NLP system, as it narrows down the focus of patent analysts onto the key technical features of patented devices by identifying sentences with greater relevance from a dynamic-static point of view. Moreover, the NLP system can be exploited to further investigate the theoretical elements of modelling constructs (e.g., State, Transitions, Events, Decision Nodes, Activities) by researchers in the fields of system modelling and linguistics and, in turn, to develop automated NLP tools capable of extracting concise and abstract representation (models) of the functioning of patented devices from text.

Appendix

In the present Appendix I provide possible improvements to the limits reported in section 5.2.3

Table 6: Limitations of the NLP system and its possible Improvements in the cases: (1) dictionary coverage and (2) overlap.

Case	Limitations	Improvements
1	Domani-specific verbs and expressions	• Expand the dictionary with field-specific expression and verbs used to express the dynamics of the technology of concern.
1	Generic expressions	• Apply text normalization to rephrase sentences with generic expressions. For instance, the expressions: *"provide sufficient processing power", "performing cultivation"* can be converted to *"power up"* and *"cultivate*", respectively.
2	Dependent Clauses	• Use the clauses of sentences as the unit of analysis (instead of sentences), thus applying a different approach on sentence splitting.
2	Polysemy of words	• Pre-process patent descriptions to remove expressions in which dynamic keywords have not their intended use such as: *"when desired", "according to the following examples", etc…*
1, 2	Sentence Splitting errors *	• Use ad-hoc Sentence Splitting modules designed for patent documents. • Filter out mis-splitted sentences by removing sentences composed of less than 10 words (too short) and sentences composed of more than 150 words (too long)
1, 2	POS-tagging errors *	• Use ad-hoc POS-tagging modules trained on patent documents. • Pre-process text to rephrase sentences with complex structure and disambiguate component names.

* Note that: the precision of an NLP system is inherently upper bounded by the precision of the POS-tagging and Sentence Splitting modules used in an NLP pipeline.

References

Cascini, G., Russo, D., & Zini, M. (2007). Computer-aided patent analysis: finding invention peculiarities. In Trends in Computer Aided Innovation: Second IFIP Working Conference on Computer Aided Innovation, October 8–9 2007, Michigan, USA (pp. 167-178). Springer US.

Chen, L., Xu, S., Zhu, L., Zhang, J., Lei, X., & Yang, G. (2020). A deep learning based method for extracting semantic information from patent documents. Scientometrics, 125, 289-312.

Fantoni, G., Apreda, R., Dell'Orletta, F., & Monge, M. (2013). Automatic extraction of function–behaviour–state information from patents. Advanced Engineering Informatics, 27(3), 317-334.

Harel, D. (1987). Statecharts: A visual formalism for complex systems. Science of computer programming, 8(3), 231-274.

Jacobson, L., & Booch, J. R. G. (2021). The unified modeling language reference manual.

Krogstie, J. (2012). Model-based development and evolution of information systems: A Quality Approach. Springer Science & Business Media.

Liang, Y., & Tan, R. (2007). A text-mining-based patent analysis in product innovative process. Trends in computer aided innovation, 89-96

Opdahl, A. L., & Sindre, G. (1995). Facet models for problem analysis. In Advanced Information Systems Engineering: 7th International Conference, CAiSE'95 Jyväskylä, Finland, June 12–16, 1995 Proceedings 7 (pp. 54-67). Springer Berlin Heidelberg

Pidd, M. (2004). Complementarity in systems modelling. Systems modelling: Theory and practice, 1, 20.