Dr hab. Inż. Maciej Piasecki
Katedra Sztucznej Inteligencji
Politechnika Wrocławska
maciej.piasecki@pwr.edu.pl

25th June 2022

Reviewer's opinion
on Ph.D. dissertation authored by
Dawid Wiśniewski
entitled:
Automating Competency Handling in Ontology Development Process

1. Problem and its impact

What is, in your opinion, the most important problem discussed in the dissertation?

The most important problem focused on in the thesis is an idea of automated procedures applied in ontology development and testing. The scope is concentrated on Competency Questions as the main source of requirements that is a good direction in comparison to methods described as 'ontology learning from texts'. Competency questions provide a good balance between the natural language as a means for specification vs necessary precision in requirements. Moreover, the problem of the automated procedures is treated in the thesis in a very comprehensive and systematic way. This culminates in the presentation of the scheme of Test-Driven Development of ontologies instantiated with different processing methods (and their implementations) proposed in the thesis.

Is it a scientific one?

The overall endeavour is very scientific in its aims and complexity. However, many of the proposed methods tend towards heuristic solving different subproblems focused on immediate practical goals. Thus, they have more engineering than scientific character and often lack some scientific rigour. Nevertheless, the thesis as a whole, and definitely not a small one whole from many angles, is an interesting scientific achievement. The signalled flaws, that will be further elaborated in the rest of this opinion, may partially result from the complexity of the thesis scope.

Does it have a practical meaning?

There are several aspects of practical importance of the work done for the thesis. A unique research dataset and several analysis tools – computer programs – have been completed and publicly published. It is of great value that all technological elements developed for the thesis are available on open licences. This well contributes to the idea of open science. Many of the proposed tools have character of proof-of-the-concept prototypes, due to their limited performance and some limitations assumed, but many of them are unique.

2. Contribution

What is the main, original contribution of the dissertation?

The author formulated (on pages 3-4) five research questions to be answered in the thesis and in relation to them the main contributions. The whole set of questions is well selected and comprises an ambitious and interesting research plan. All the five research questions have been approached and quite advanced results have been obtained. The main original contribution of the thesis is showing an example implementation of the idea of Test-Driven Development of ontologies focused on competency questions as the main source of requirements. Secondly, a unique and good scientific resource "CQs translated into SPARQL-OWL" has been designed, built and published with some minor issues related to its quality evaluation. The other elements of the whole process, namely "automatic method for recommending translations of ontology competency questions into SPARQL-OWL" (Chapter 9) and "the method of generating synthetic pairs of CQs and SPARQL-OWL queries" (Chapter 8) described as contributions are less scientifically advanced, nevertheless completed as fully functional proofs of the concepts. It means that complete methods and implementations were provided for these subproblems, but not necessarily the best possible ones in relation to the state of the art. Arguments for this opinion are given in Point 3 below.

3. Correctness

Can we trust what is claimed in the dissertation?

The dissertation is written in a very systematic and careful way. All the decisions and steps are very well described.

Are the argument s correct? Indicate the flaws you have noticed, if any.

The main recurring problem is that the work in dissertations encompasses several subdomains, that would require separate overviews of the state of the art, and the proposed solutions, mostly heuristic, could benefit from detailed analysis of the literature and solutions in these subdomains.
The dissertation includes a very good introduction to the basic notions and methods in Machine Learning with carefully prepared illustrations. However, it expresses some imbalance: elementary notions received more attention, like the basics of neural network, while methods of practical importance, like recurrent neural networks or transformers, are discussed only shortly. What is more, the overview of the state of the art for the issues in Chapters 7 – 9 is done very selectively. Even the notion of presupposition – the key one for Chapter 10 – is rather mentioned than properly discussed.
A reverse division of attention would be more feasible: less effort paid to text-book-like part, more to the comparison with the state-of-the-art.

The dataset, scientific resource, of "CQs translated into SPARQL-OWL queries" is one of the valuable achievements of the dissertation. However, its evaluation has not been shown in all details. First of all, the translations were done by single experts and the inter-annotator agreement, or a similar measure, was not calculated. The number of potential inconsistencies and translator idiosyncrasies is not known. This is a recurring problem in the dissertation. Several tasks in manual preparation of some datasets were done by single persons per dataset element. In addition, many CQs could not be translated that raises questions about the coverage. Once again, such an attitude re-appears in the dissertation, some limiting assumptions are introduced, that results in limited coverage. In several cases existing gaps were next identified in analysis, but they stay and their influence on potential applications is not estimated.

Concerning the tagger in Chapter 7, with so limited and synthesised dataset, the idea of trying rule-based solution is very good. The achieved results are promising. The only drawback is the development of the whole tool from scratch, while rule-based technology in Information Extraction has long and still live tradition. Considering the expressive power of the proposed language of rules, it would have been good to consider using an Information Extraction platform like GATE. This could facilitate the reusability and reproducibility of the obtain results. For instance, GATE provides modularity and different control strategies besides many other useful features. A lot of work done in the case of the rule-based extractor seem to be redundant to the existing IE frameworks.

The proposed construction of the tagger (extractor) based on Machine Learning follows some not convincing decisions. Besides the lack of use of vector-based text representation (embeddings), that was commented later in the dissertation by the author itself, the training set was somehow generated on the basis of simple patterns that may hamper the tagger. Contrary to this low diversity of the training set, the number of features is very high in relation to the dataset size, but also diversity. It is good that parameter optimisation was applied, but it would be also very useful to perform feature selection, that was omitted.

"Overlap resolver", contrary to its name, deals only with the cases of inclusion and improper overlapping (identity of sequences). What did happen with the case of genuine overlapping, i.e. partial? It would be strange not to see such cases.

In the case of the evaluation of both taggers, it was a very good decision that a separate set of ontologies was selected for the test set. It is a little bit unclear why does the validation set contain all other ontologies than SWO and the proper training set include only SWO? It would be good to have more diversified composition of the training set. The assumed one may result in a negative bias. Moreover, proper split of data on the level of concepts and terms could be done horizontally, across the ontologies.
In NLP, especially in relation to the evaluation of taggers and NERs (Named Entity Recognition tools), the notions of strict and weak accuracy (or other similar measures) were introduced, but there are no references to them in the dissertation. Moreover, it would be also helpful (and is a common practice in NLP) to measure quality of boundary detections and full extraction, as two consecutive analysis levels (even if done in one gone, that is completely correct).

In Chapter 8, "the method is to generate a large set of diverse CQ forms mapped to SPARQL-OWL queries targeting the most prevalent modelling decisions." is a relatively straightforward but innovative and valuable work. However, several limiting assumptions were introduced, like: "For this reason, we decided to make our method able to generate questions asking for at most a single thing." (page 91). What is worst, such simplifying assumptions only add to the limitations stemming from a controlled, semi-formal language used as a basis, e.g. in automated generation.
The work in Chapter 8 and 9 is not informed by or compared to quite rich tradition of rule-based translation systems or semantic parsers (in fact translating text to formal representations), as well as Natural Language Generation systems. If such a study had been done, the work of the author would have been much easier. What is more, many solutions worked out in the dissertation, have a heuristic character, are meticulously constructed from scratch, hard to be generalised and further expanded.

On page 94: "For this reason, we decided to strip the word every when filling placeholders in CQ templates." — this seems to be a dramatic intervention into the semantics! However, fortunately,

somehow, it is legitimate due to neglecting many formal issues in Chapter 8. A similar semi-formal approach was applied to presupposition in Chapter 10, as it is discussed below.

There are different language resources describing synonyms, many of them are open. The author built his own list without providing arguments for not using or even comparing to the existing lexicons of synonyms. The notion of synonymy, a difficult one, has not been characterised in the dissertation. When we take a look into the list in the appendix, we can notice that a) it is very limited, and b) the synonyms have a heuristic character, focused on particular use in the author's method. Concerning the latter, it may be accepted, but an operationalised definition of synonymy is necessary, also for expanding the list in the future.

Coming to evaluation in Chapter 8, on page 95 we find "we generated a comprehensive list of CQ templates," the question is how was this comprehensiveness verified? The coverage reported in Table 8.8 may look quite limited (it could be expected remembering many simplifications signalled in the method description). So, one may have a question for what kind of applications is this enough? It may be sufficient, probably is sufficient, but it would be good to analyse and show this in the dissertation.
In the report on gaps in coverage, we may find "page 99: "is filled with a noun in plural, the word is should be replaced with are and a should be omitted." so what is a problem in having this done? It seems that here the heuristic and engineering character of the proposed algorithms for translation finds its natural limitation preventing further expansion – it becomes too complicated and amorphous. Casting the problems in terms of a rule-based Machine Translation system or a semantic parser, with clear distinction between knowledge resources, including lexicons, grammars, etc., would facilitate tackling different detailed issues, and could introduce more clarity into the structure of the methods.

Chapter 9 deals also with a translation (or semantic parsing problem) and starts with surprising revisiting of the tagging problems. This time, firstly, from the lexicon-based tagger. However, it is worth to emphasise that the author is aware of the lack of consistency and coordination between Chapter 7 and 9, explains this by the chronological order of the work on dissertation, and this is quite understandable. Finally, ReqTagger has been used as an element of the processing in Chapter 9. It is worth to notice that the author used in Chapter 9 contemporary method for text representation, namely contextual embeddings in Step 5 (Phrase linking) and presented good motivation for this showing his good understanding of this kind of methods. In some contrast to this, in Step 3 (Closest known CQ pattern selection) of the same method word n-grams (n element sequencies) are used for comparing CQ pattern, while such patterns express intrinsic graph structure that could be leveraged for this purpose. In a similar way to other methods in the dissertation, several limiting assumptions are introduced, e.g., "We assume that each predicate chunk PC representing a property in a SPARQL-OWL query template must have exactly two arguments". Thus, page 110, "that SeeQuery provides correct recommendations for 46 out of 62 CQs (74%)" - this is significantly below 100%, could be valuable, some argumentation in favour of this coverage level is missing.
No arguments are also given for the selected score values. Evaluation is also based on decisions of a single expert, so any quality factors cannot be computed.

Chapter 10 starts with a very good idea of using the notion of presupposition in the context of a system interpreting queries. It is not a novel idea in the case of dialogue systems or, in a broader perspective, of systems of semantic analysis of statements in natural language (e.g., the Polish system Polint built by Zygmunt Vetulani in 90s). Such approaches are very close to the aims of the system in Chapter 10, so it would have been good to compare to and base on them. Questions induce existential presupposition as it was noticed and used by the author. Depending on the type of question particle

situation is more complicated, but just concentrating on simple existential presuppositions is an acceptable technical simplification. However, unfortunately for the formal clarity of the method, the author immediately starts heuristically expanding the mechanism of presupposition with additional techniques that could be called informally `negative presupposition'. For instance, on page 114: "the presupposition can be denied" – if it can be, it is NOT a presupposition! Presupposition cannot be denied, this is a fundamental property of presuppositions! If we go so far from presuppositions, we should stop calling them presuppositions! Moreover, on page 115 we find: "two kinds of presuppositions can be identified that both should be satisfied to answer the question" – this is simply not true, "which software" introduces the existential presupposition related to "software" independently of its properties!
The idea of 'negative presupposition' may be useful from the engineering point of view, but it is formally incorrect. It would be better to firmly base considerations on the formal ground and clearly mark heuristic solutions inspired by the notion of presupposition as a separate technique.
Fortunately, Chapter 10 includes a very good interpretation of the test-driven ontology authoring that was elaborated with solutions proposed in the thesis. The author presents in a convincing way how the whole process can be implemented with the proof-of-the-concept solution he presented in the subsequent chapters of his dissertation. From this perspective, the assumed shallower, but wider research plan starts to show its positive aspects.

4. Knowledge of the candidate

What are the chapters of the dissertation (or sections in chapters) that resemble a tutorial and thus confirm a general knowledge of the candidate in the discipline of ***Information and Communication Technology****. What areas of that discipline are covered by those chapters/sections? What do you think about quality of those chapters/sections?*

As it was already noticed, the first part of the dissertation is a very good and systematic introduction to basic notions and methods in ontologies, Machine Learning and Natural Language Processing. This part is very well written and pleasant in reading. Thus, the author showed his good and broad knowledge in Information and Communication Technology.

What is your opinion on the list of references? What is the degree of its completeness?

On the general level the list of references is a good background for the discussed issues. As it was already mentioned, many works and solutions related to different tasks and subdomains pertaining to the different chapters have been overlooked. The wide range of the works planned for the dissertation could have caused that the author too quickly followed his own intuitions without comparison to the literature, but these problems have been already discussed.

5. Other remarks

Additional detailed comments:

p 27: " assign labels spanning over multiple elements in a sequence when solving structured prediction problems. Hereafter, we refer to such spans as entities" — not every sequence of "elements" is "an entity"; what is more there are no entities in text, but only language expression that can represent entities

p 27: "elements of sequences that form entities " — text elements do not form entities, unless we are talking about some language (linguistic) entities

p 30: "Humans think in terms of concepts and relations between them, and these are linked to words, no st the whole texts." — this is not true or much too far going simplification! E.g. concepts may be even implicitly expressed in an utterance

p 30: "a list of meaningful elements: words, punctuation," — this is not a full list of tokens, moreover, what a token is depends on a model

p 34: "4.4 Regular expressions" one would expect to see this section earlier, e.g. before parsing at least

p 35: "sum of vectors assigned to the n-grams " — fastText, the sum?

p 35: "The bidirectional LSTM consists of two LSTM cells" — two types of cells?

p 41: "the masked sequence is constructed" — masked or marked? marked in the previous sentence

p 62: "Create a Cartesian product between all anygrams created from CQs" — a little bit imprecise, but the idea can be guessed

p 77: "We handcrafted a list of rejected phrases based on the training set and our expertise." — typically such a list is called a stop list

p. 79: "However, although the CRF-based tagger outperforms ReqTagger in terms of precision, it has a very low recall." — this is very strange and counter-intuitive, it should be the other way round (ML based methods usually have better recall than rules). As there are no miracles, but only bugs in science, the observed results is probably caused by too many features and too small data, or some lack of simple post processing of the results of the ML-based tagger.

p 80: "the whole Which software is extracted as an entity suggestion." — but it could be very simply corrected! using ML does not mean that any kind of rule-based post processing is forbidden

p 81: "the POS tagger makes wrong decisions assigning a verb (VBZ) to a noun gateways." — it would be interesting to measure and know the PoS tagger performance on this particular dataset

p. 81: "If there is a determiner before the verb, a new rule …" — the is not a verb, but a past participle, "matching a verb in the past" — wrong categorisation

p. 81: "than noun phrases defined as in spaCy." — a mental shortcut, NPs extracted by a tool from spaCy

p. 83: "shows that CQ2SPARQLOWL does not introduce some CQ forms" — introduce? or rather includes or shows?

6. Conclusion

Taking into account what I have presented above and the requirements imposed by Article 187 of the A*ct of 20 July 2018 - The Law on Higher Education and Science* (with amendments)[1], my evaluation of the dissertation according to the three basic criteria is the following:

A. Does the dissertation present an original solution to a scientific problem? (the selected option is marked with **X**)

☐	☒	☐	☐	☐
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

B. After reading the dissertation, would you agree that the candidate has general theoretical knowledge and understanding of the discipline of **Information and Communication Technology**, and particularly the area of automated ontology development?

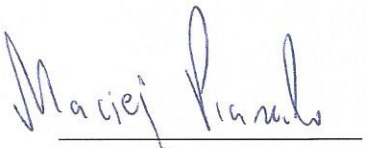
☐	☒	☐	☐	☐

[1] http://isap.sejm.gov.pl/isap.nsf/DocDetails.xsp?id=WDU20190000276

Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

C. Does the dissertation support the claim that the candidate is able to conduct scientific work?

[X]	[ ]	[ ]	[ ]	[ ]
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*



Signature