Knowledge Graph and Natural Language Processing to support Engineering Design

Appendix B. Evaluation of the KG dimensions: quantitative and qualitative dimensions

In this Appendix we describe specifically the comparative evaluation of quantitative dimensions (Appendix B.1) and quantitative dimensions (Appendix B.2). We start from describing the available candidate benchmarks for the evaluation.

First, Siddharth et al. (2021a) proposed a NLP approach to build a large-scale engineering KG from patents. The core of their work is the extraction of triples from patents. However, this valuable approach is not compatible for an evaluation with TecKnoGraph for two reasons: (i) no open data is shared to access the results of the work; (ii) no classification of the entities and relations is performed. This latest issue makes this work incomparable because TecKnoGraph goes beyond the mere extraction of triples.

Second, Jang and Yoon (2021) proposed TechWordNet, a semantic relation network of technologies built using NLP and deep learning techniques. This work is also not compatible with TecKnoGraph for two reasons: (i) no availability of their data resulted from the experiment; (ii) it focuses on building a KG only considering *technologies*. While the focus of our work is to find novel relations among more than one entity, their work focuses more on the extraction method. Even though they also build a KG, it is impossible to compare it with TecKnoGraph because they do not provide a comprehensive assessment of the quality of the results.

This lack of benchmarks leads us to perform an evaluation of the resulting KG from our experiment by comparing it with a variant. We propose a variant that is based on common-sense knowledge. In particular, we decide to evaluate the resulting dimension quality of the application TecKnoGraph on C4ISTAR with a KG whose classification of relations relies on ConceptNet.

Let us consider the following two KGs:

- ***tKG**: TecKnoGraph*. This KG is the result of the experiment proposed in this paper.
- ***tKG-c**: TecKnoGraph with ConceptNet*. This KG is the result of the experiment proposed in this paper but with a difference in the Classification phase.

tKG-c differs from *tKG* in the method of classification of relations. While *tKG* classifies relations with a bottom-up approach using word2vec (see Section 3.2.3), *tKG-c* classifies relations according to the ConceptNet[1] *supersenses* (Speer et al., 2017). *Supersenses* are coarse-grained semantic categories associated with each word (Flekova and Gurevych 2016). In *tKG-c,* given the verb lemma of each relation we classified it using the ConceptNet API. Given that a relation always contains a verb, we classify the relation according to the supersense of its verb. For example, given the triple <operator;move;gripper>, the relation is classified as "*motion*" because its supersense is "*motion*". However, given the polysemy of language, ConceptNet provides multiple supersenses for each term. For example, the verb "*move*" could be classified with "*motion*", "*contact*" or "*communication*".

The main difference between the two KGs is that the classification of their relations relies on two different approaches: *tKG* is bottom-up while *tKG-c* is top-down. The use of ConceptNet as a knowledge base to classify relations has several limitations. Given that it is not domain-specific, ConceptNet attempts to map knowledge in the broadest possible way (Speer et al., 2017). For example, while ConceptNet aims to keep information about language polysemy, ED knowledge bases aim at the opposite: being able to remove ambiguity of language (Fantoni et al., 2013).

This problem emerges when analyzing two particular dimensions of the obtained KG: consistency and completeness. Consistency, defined as the degree to which the knowledge of a KG does not contradict itself (Zaveri et al., 2013), is higher in *tKG* because the probability of finding dispersed distribution of relations is lower than in *tKG-c.* Completeness, the extent to which a KG covers the knowledge of interest (Stvilia et al.,

[1] We used ConceptNet version 5: https://conceptnet.io/

2007), is higher in *tKG* as a natural consequence of higher consistency (and vice versa, of the inconsistency of *tKG-c*).

A detailed description of the calculation of the evaluation of quantitative dimensions (accuracy and consistency) and qualitative dimensions (timeliness, trustworthiness, availability and completeness) is provided in Appendix B. (i) The evaluation of the accuracy is not able to discriminate which is the KG with better quality; this leads us to consider that tKG-c is more inconsistent than tKG (see Appendix B for more details). (ii) trustworthiness is equal for all the approaches; (iii) the availability does not discriminate against the quality of the compared KGs, given that we are comparing two variants of the same KG. (iv) The completeness is higher for *tKG* because ConceptNet introduces misconceptions due to the language polysemy (Fantoni et al., 2013). The overall comparative evaluation shows that tKG outperforms tKG-c in terms of consistency and completeness.

B.1 Evaluation of quantitative dimensions

We can now discuss the single evaluation metrics of the KG, in particular, focusing on a comparative analysis between the tKG and the tKG-c. These dimensions can be divided into qualitative and quantitative. We now discuss the evaluation of the quantitative dimensions: *accuracy* and *consistency*.

The *accuracy* refers to the degree of correctness of facts in a KG (Zaveri et al., 2013), namely “the extent to which knowledge is correct, reliable, and certified free of error” (Wang & Strong, 1996). The accuracy of our comparing KGs regards the phases of Extraction and Filtering. Given that *tKG* and *tKG-c* differ only for the Classification phase, the evaluation of the *accuracy* is not able to discriminate which is the KG with better quality.

The *consistency* is defined as the degree to which the knowledge of a KG does not contradict itself (Zaveri et al., 2013). Contrary to *accuracy*, *consistency* focuses on the logical correctness of a KG rather than the semantic correctness of triples. This metric is hard to assess manually (Ulbricht et al., 2018; Wang et al., 2021). This is due to the large number of entities and relations present in a large-scale KG. However, Haase and Völker (2008) states that *inconsistency* of a KG “is correlated with the probability of finding dispersed distributions of relations”. This means that a KG with high dispersion of relations has a high probability to be inconsistent. In practice, inconsistency is proportional to the dispersion of relations in the KG (Zhang et al., 2021). In our study, given that we are comparing two KGs that differ in their relations classification, we can assert the following statement: the KG with higher dispersion of relations has higher probability of containing inconsistency. Hence, we proceed in the calculation of the relations dispersions of both *tKG* and *tKG-c*.

Given that both KGs present a structure where relations are classified according to clusters aggregated from multiple entity-pairs, the calculation of the dispersion needs to be performed at different levels (Burgin and de Vey Mestdagh, 2015). In fact, to compare the two KGs, we employ the dispersion formula $D = \sigma^2/\mu$ and we apply it at (i) relations level, (ii) at clusters level, (iii) at entity-pair level and then at (iv) KG level. We proceed as follows.

For what concerns the dispersion at relations level we need to consider that a general relation r_i can be found with certain frequencies between different entity-pairs. For example, the relation “*consist_of*” can be found with a certain frequency between two *technologies* and with a certain frequency also between a *technology* and an *advantage*. For this reason, a general relation r_i can be expressed as a vector π_i of frequencies between all the combinations of entities. This vector π_i expresses the distribution of the relations among the possible entity pairs. This vector does not differ between the *tKG* and *tKG-c*. Different relations have different vectors, and consequently, different distributions. In fact, relations that appear in a lot of entity pairs are less concentrated (or more dispersed) than relations that appear in only one entity pair. The dispersion at relations level is calculated as the $D_r = \sigma_r^2/\mu_r$ of the vectors π_i.

For what concerns the dispersion at cluster level we can start to consider c_i as a cluster of relations extracted from two entities. We can represent c_i as a n-dimensional vector whose elements are the dispersion indexes of each relation $\{D_r, \dots , D_n\}$, where n is the number of relations in the cluster. The dispersion of the clusters is given by the dispersion index $D_c = \sigma_c^2/\mu_c$ of the elements in c_i.

For what concerns the dispersion at entity-pair level we can consider p_i as the set of the dispersions of all the clusters occurring in an entity-pair. We can represent p_i as a m-dimensional vector whose elements are the dispersion indexes of each cluster $\{D_c, \dots , D_m\}$, where m is the number of clusters in a given entity-pair. The dispersion at entity-pair level is given by the dispersion index $D_p = \sigma_p^2/\mu_p$ of the elements in p_i.

Finally, for what concerns the dispersion at KG level we can consider g_i as the set of dispersions of all the entity-pairs present in the KG. We can represent g_i as a k-dimensional vector whose elements are the dispersion indexes of each entity-pairs in the KG $\{D_p, \dots , D_m\}$, where k is the number of entity-pairs present in a KG. In this case m is equal for both *tKG* and *tKG-c*. The overall dispersion of the two KG is finally determined by the dispersion index $D_g = \sigma_g^2/\mu_g$ of the elements in g_i.

	tKG *- TecKnoGraph*	***tKG-c*** *- TecKnoGraph with ConceptNet*
Relations dispersion	0.33	0.33
Clusters dispersion	0.23	0.34
Entity-pairs dispersion	0.21	0.44
KG dispersion	0.20	0.38

Table B.1 - Results from the comparative evaluation of inconsistency between tKG and tKG-c. This table shows the dispersion index of (i) relations, (ii) clusters, (iii) entity-pairs and the (iv) overall dispersion of the two graphs.

Table B.1 shows the results dispersion calculation comparing the two *tKG* and *tKG-c*. As we can see, the dispersion at cluster, entity-pair and KG level is higher for the KG built with ConceptNet. This leads us to consider that *tKG-c* is more inconsistent than *tKG*. High dispersions lead to high probability of finding inconsistent facts inside a large KG (Heyvaert et al., 2019). This would be the case of TecKnoGraph whose relations are classified according to common-sense knowledge.

B.2 Evaluation of qualitative dimensions

We now discuss the evaluation of the qualitative dimensions: *timeliness, trustworthiness*, *availability* and *completeness*.

The *timeliness* is defined as the degree to which knowledge is up-to-date. For example, a daily updated KG can be considered to have high timeliness. Both benchmarks analyzed patents until the date of publication of the respective papers. The present study performs analysis until the same date of the benchmark (year 2021). For this reason, neither of the KGs outperform in terms of *timeliness*.

The *trustworthiness* is the degree of objectivity, authority, and verifiability of a KG. It describes how credible the knowledge in a KG is accepted to be. Trustworthiness is a comprehensive dimension with a subjective perspective, such as authority of knowledge. Given that *trustworthiness* can be assessed through the *trustworthiness* of knowledge sources (Wang et al., 2021), we can state that this dimension is equal for all the approaches.

The *availability* is the extent to which knowledge is convenient to use. Wang and Strong (1996) refers to it as "how easily and quickly the knowledge of KGs can be retrieved". Given that we are comparing two variants of the same KG, the *availability* does not discriminate the quality of the compared KGs. However, as stated in Section 3.4, we provide open access to the code and to a piece of KG extracted from the experiment.

The *completeness* refers to the extent to which a KG covers the knowledge of interest (Stvilia et al., 2007). Generally, a KG is considered to contain good completeness if it meets the knowledge interests of the underlying applications. Different methods have been proposed to evaluate KG completeness. However, Mendes et al. (2012) provide a valuable insight on this measure: the *completeness* of KGs strongly depends on the approach employed in the construction phase. In particular, Mendes et al. (2012) demonstrated that a KG built using a bottom-up approach outperforms a KG built with a top-down approach. The distinction of these two approaches fits the distinction between *tKG* and *tKG-c*. In fact, while both KGs are automatically extracted from texts, *tKG* classifies relations with a bottom-up approach but *tKG-c* classifies relations with a top-down approach. Namely, in the Classification phase, *tKG-c* relies on an external pre-existing knowledge base. The use of general purpose knowledge base makes the resulting *tKG-c* less complete than *tKG.*

A more detailed discussion on the *completeness* could be given if we inspect the results of the Classification phase of *tKG-c* and we compare it with the results of the *tKG.*

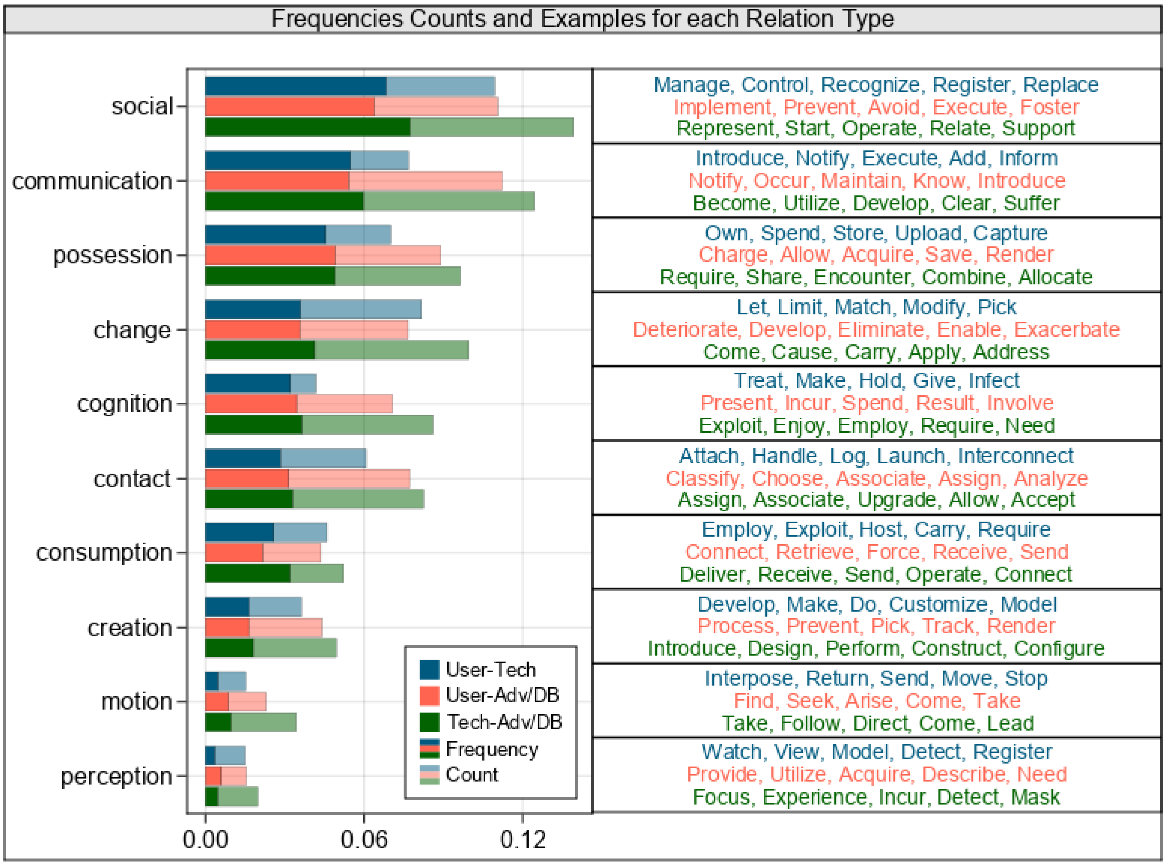

Figure B1 - This figure represents a study of clusters for relations linking users and technologies. ***On the left:*** *the count and the frequency of words, both summing to one over the whole set, for each cluster.* ***On the right:*** *the 10 most frequent elements of each cluster. Dark blue is the frequency of the entity pair "user-tech". Light blue is the count of the entity pair "user-tech". Dark red is the sum of the frequency of the entity pair "user-advantages" and "user-drawbacks". Light red is the sum of the count of the entity pair "user-advantages" and "user-drawbacks". Dark green is the sum of the frequency of the entity pair "technology-advantages" and "technology-drawbacks". Light green is the sum of the count of the entity pair "technology-advantages" and "technology-drawbacks"*

Figure 5 presents the ranking of the top 10 classes. In particular, in this plot the *frequency* and the *count* are reported. The former is the measure of occurrence of the verbs assigned to a given *sense* in the patent set analysed within the scope of this work. The latter is the measure of the size of the *sense*, and so the number of extracted verbs related to a given *sense*. We chose the *frequency* to define the rank presented in Figure 4 because it explains the importance of a given class in the examined documents, while the *count* depends on the size of the sense in the ConceptNet itself.

Given that these categories are "a discrete representation of one aspect of the meaning of a word" (Jurafsky and Martin, 2000), we need to critically discuss our results. Common sense knowledge is wider than design knowledge and it comprises several problems due to the polysemy of language (Fantoni et al., 2013). Consequently, the fact that we use common sense knowledge to capture and represent design knowledge can lead to misleading insights. These misconceptions are strong evidence of a critical lack of *completeness* of the *tKG-c*.

For example, *use* and *utilize* are not in the same categories. Although they are synonyms, it is critical to find them in two different classes. Other criticalities come from verbs such as *hold* and *search* that are present respectively in the "*Creation*" and *"Contact"* categories. In technical language, holding something does not imply creation, as well as searching for something does not imply being in contact. All of these wrong associations are the consequence of the fact that language has a large polysemy.

The overall comparative evaluation shows that *tKG* outperforms *tKG-c* in terms of *consistency* and *completeness.* We show that the construction of a KG with the use of an external common-sense knowledge base can lead to inconsistent and incomplete knowledge if compared to a KG constructed exploiting domain knowledge.