1 STUDY 4: EVALUATING EACH STEP OF THE METHODOLOGY

We evaluated in Study 1 and 2 the performance of our approach, in which the results are mainly based on the performance of the defined rules (step 4). In Study 3, we evaluated the impact of the lightweight representation on ideation outcome. In this fourth study, we investigated the performance of the remaining steps of the proposed solution i.e. identification of ideas' elements (a single term, sequence of words, or triple) and matching mechanism, in order to have a better understanding of the results obtained in Study 1 and 2. Therefore, we first determined the performance of the identification of ideas' elements and then the results of each matching technique.

1.1 Measures

We created two ground truths, one for ideas' elements and another one for correct correspondences between terms and predicates pairs. Then we compared the identified elements and the correspondences found by our approach to this ground truth.

Due to time constraints, we took 1/3 of data and randomly selected 35 ideas of 95 ideas generated in the first study[1]. Next, two of the researcher experts have identified and validated the ideas' elements and correspondences in a four-round process. In the first round, one expert identified different elements e.g. which one should be a single term, sequence of words, or triple in patterns WHAT and HOW, while the second expert was focused on defining the correspondences between terms and predicates pairs. In the second round, the elements identified by the first expert were given to the second expert for validation and the established correspondences by the second expert were given to the first expert for validation, we computed Cohen's kappahttps://idostatistics.com/cohen-kappa-free-calculator/ and we report coefficient of 0.69%. This coefficient is considered as substantial agreement which due to lack of context. For instance, the term display can be either "Graphics device" or "Information display". In the third round, both experts come together and discussed the ambiguous cases to generate a shared understanding. In the final round, both experts coded the remaining terms and made a final check to see if some cases need adaptation.

Next, we measured *precision*, *recall* and *F-measure* for determining the quality of both steps (identification of elements and matching mechanism). *Precision* is the number of correctly 'identified elements' relative to the number of all extraction carried out. *Recall* defines the number of correctly 'identified elements' relative to the number of all correct 'identified elements'. The *F-measure* is the harmonic mean of precision and recall. We replaced 'identified elements' with 'established correspondences between subject/object or predicate pairs', then we applied the same measures to evaluate the matching mechanism.

1.2 Results

Pattern	#Ideas	Idea Ele.	Prec.	Rec.	F-meas
WHAT	35	Triple	0.818	0.277	0.40
		SW	0.60	1	0.75
		single term	1	0.4	0.571
HOW	35	Triple	0.727	0.307	0.432
		SW	0.227	0.833	0.357
		single term	/	/	/

Table 1: Results of the Identified Ideas' Elements single term, triple or sequence of words (SW). The symbol "/" in the table means that no value was calculated since the ideas' mechanism where not describe in single term

1.2.1 *Identified Elements.*

Table 1 summarises the results obtained when processing the patterns WHAT and HOW of the ideas. The second column indicates the number of ideas considered for the evaluation. The users describe the patterns of their ideas either in single term or sequence of words. Thus, the system attempted to extract triples only from those patterns described using a sequence of words. The row Triple refers to cases when the patterns where expressed in sequences of words and it was possible to extract triples from them. For other cases sequence of words(SW) and compound(W) it was not possible to extract triples. The third column presents different elements identified in patterns WHAT and HOW. Those identified elements are then compared to the ground truth and the remaining columns show the *Precision*, *Recall* and *F-measure* for each pattern. Our results can be summarised as follows:

- The recall values were low for single term detection and much lower for triple extraction
- In contrast to the pattern HOW, the pattern what can be described in single term (e.g. emergency TV)

1.2.2 *Matching Mechanism.*

We conducted a detailed evaluation of the ideas' elements matching mechanism. The matching mechanism receives as input either single term or sequence of words and generates as an output a relation between them, where possible types of relation are Equals, Synonym, Similar, Disjoint, Subclass/Superclass. For the first case, a single term can be either a compound noun or a noun (subject/object of a triple, or a single term used to describe the whole pattern) or a verb (predicate of a triple). For the second case, a sequence of words is considered when no triple has been extracted.

[1] We provide this data set for other researchers at:
https://github.com/PaperMaterial/submission905

Setting Up the Threshold The application of a threshold is a standard technique in matching systems. Two entities with a similarity value greater than a given a threshold are considered as correct i.e. similar otherwise they are deemed incorrect i.e. disjoint. We evaluated our approach with thresholds ranging from 0.55 to 0.95. For each technique, we computed the F-measure, which is the harmonic mean of precision and recall, and computed the mean of this value over all pairs of entities of 35 ideas. Based on results obtained (see Figure 1 and Table 2), we fixed the threshold for each matcher.

Thr.	**0.6**	**0.65**	**0.70**	**0.75**	**0.80**	**0.85**	**0.90**
SM	0.554	0.554	0.554	0.554	0.554	0.554	0.554
LM	0.419	0.419	0.445	**0.476**	0.476	0.476	0.476
CM	0.332	0.435	0.460	0.531	**0.570**	0.557	0.559
TM	0.414	0.480	0.519	0.600	0.646	**0.676**	0.607

Table 2: F-measure values for each matcher averaged over 35 ideas' terms for thresholds ranging from 0.50 to 0.90. The results of structure-based matcher (SM) do not depend on threshold value because the relation between two terms either is defined in Wikidata or not; the linguistic-based matcher (LM) performs best for a threshold 0.75; the content-based matcher (CM) performs best for a threshold 0.80; Terminology-based matcher (TM) performs best for a threshold 0.85.

Performance of Matchers The tables 3a, 3b, 3c summarise the outcome of each technique used in the matching mechanism for all cases: compound nouns or a noun, verb and sequence of words, respectively. We can observe that all techniques were applied and the results were very positive in terms of precision exceeding 0.8 almost all cases, although the recall values were lower, especially for the techniques that are based on DBpedia/Wikidata and WordNet.

	# NNs	**Prec.**	**Rec.**	**F-meas.**
Structure-based	66	1	0.22	0.36
Linguistic-based		1	0.25	0.4
Content-based		0.9	0.48	0.63
Terminology-based		1	0.66	0.79

(a) Results of Matching Mechanism for compound nouns or a noun (NNs)

	# VB	**Prec.**	**Rec.**	**F-meas.**
Structure-based	45	0.75	0.33	0.46
Linguistic-based		1	0.34	0.50
Content-based		1	0.2	0.33
Terminology-based		/	/	/

(b) Results of Matching Mechanism on Predicate (VB: Verbs). The symbol "/" in the table means that no correspondences were found using Terminological-based technique

	# SW	**Prec.**	**Rec.**	**F-meas.**
Structure-based	36	/	/	/
Linguistic-based		/	/	/
Content-based		0.8	0.44	0.57
Terminology-based		1	0.5	0.67

(c) Results of Matching Mechanism on sequence of words (SW). The symbol "/" in the table means that Structure and Linguistic-based techniques were not applied since the input is a sequence of words

Table 3

We have analyzed the reason for the low recall values for the techniques that are based on DBpedia/Wikidata and WordNet. We hypothesized that some entities (compound nouns or a noun and verbs) are not defined in either DBpedia/Wikidata and WordNet. To validate this hypothesis we used DBpedia spotlight to validate the performance of structure-based matcher by checking whether terms exist in DBpedia.

Results in tables 4a, 4b affirm the hypothesis and indicate a low recall for terms and verbs that have been found in DBpedia/Wikidata and WordNet.

1.3 Discussing Results of Study 4

The results show that both the identification of ideas' elements and the matching mechanism steps impacted on the identification of relationships between ideas' descriptions.

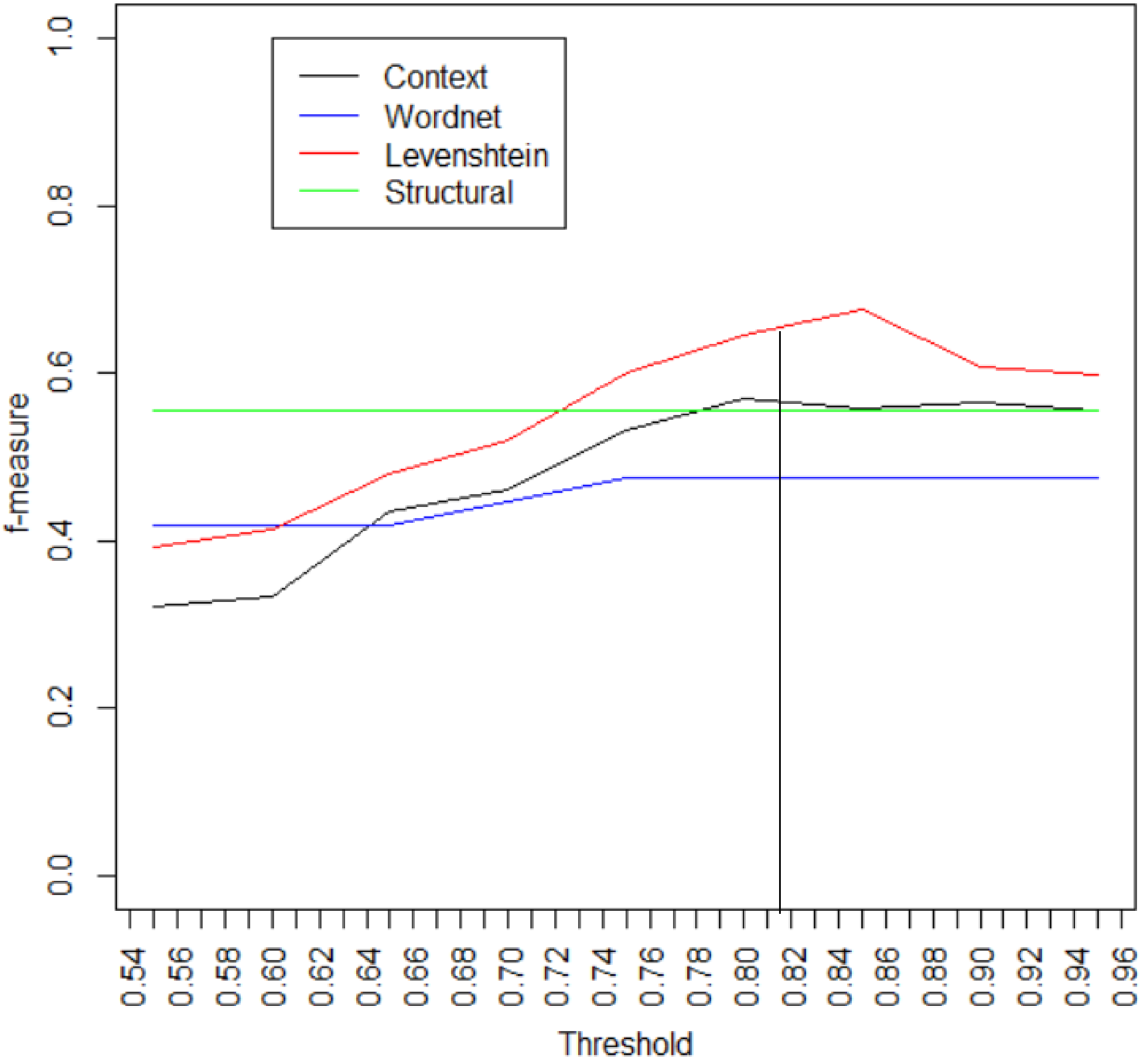

Figure 1: F-measure values for each matcher averaged over 45 ideas' terms for thresholds ranging from 0.55 to 0.95.

First, the extracting triples mainly struggled when relations are not expressed as a verb phrase followed by a simple noun phrase. This explains the low value of recall. On the other hand, the precision exceeds 90%. However, the precision did not reach 100% since some triple extractions were incoherent (e.g. structure, opens, person). This is due to the model which makes a sequence of decisions about whether to include each word in the relation phrase. A possible solution to address this problem is by considering the argument structure of the lexical items.

Second, the results of the techniques based on DBpedia/Wikidata and WordNet struggled in establishing more correspondences between ideas' entities. The reason behind it is that half of the entities were not found in either knowledge graph even if using DBpedia spotlight with confidence=0.

Our investigation shows users express their ideas by combining existing concepts e.g. emergency TV, exit path, variable structure which usually do not exist in knowledge graphs. A possible solution to cover this issue is by linking an entity to one of its generalisations. Finally, the content-based technique was not better in terms of performance. A possible explanation is that the training corpus used for our problem is not specific for this domain. A possible solution to cover this is to train a model on ideas that are generated for a specific problem e.g. urban security.

	# NNs	Prec.	Rec.	F-meas.
DBpedia spotlight	44	0.904	0.454	0.605
Search WordNet	44	0.947	0.409	0.571

(a) Results of Term i.e. compound nouns or a noun (NNs) Recognition in DBpedia/Wikidta and WordNet

	# Verb	Prec.	Rec.	F-meas.
Query Wikidata	19	0.86	0.46	0.59
Searching WordNet	19	0.91	0.68	0.77

(b) Results of Predicate Recognition in DBpedia/Wikidta and WordNet

Table 4