Knowledge-based Sense Disambiguation of Multiword Expressions in Requirements Documents

Tobias Hey*, Jan Keim†, Walter F. Tichy*
Karlsruhe Institute of Technology (KIT)
*Institute for Program Structures and Data Organization
†Competence Center for Applied Security Technology (KASTEL)
Karlsruhe, Germany
hey@kit.edu, jan.keim@kit.edu, tichy@kit.edu

***Abstract*—Understanding the meaning and the senses of expressions is essential to analyze natural language requirements. Disambiguation of expressions in their context is needed to prevent misinterpretations. Current knowledge-based disambiguation approaches only focus on senses of single words and miss out on linking the shared meaning of expressions consisting of multiple words. As these expressions are common in requirements, we propose a sense disambiguation approach that is able to detect and disambiguate multiword expressions.**

We use a two-tiered approach to be able to use different techniques for detection and disambiguation. Initially, a conditional random field detects multiword expressions. Afterwards, the approach disambiguates these expressions and retrieves the corresponding senses using a knowledge-based approach. The knowledge-based approach has the benefit that only the knowledge base has to be exchanged to adapt the approach to new domains and knowledge.

Our approach is able to detect multiword expressions with an F_1-score of 88.4% in an evaluation on 997 requirement sentences. The sense disambiguation achieves up to 57% F_1-score.

***Index Terms*—Multiword Expressions, Word Sense Disambiguation, Requirements Engineering, Natural Language Processing**

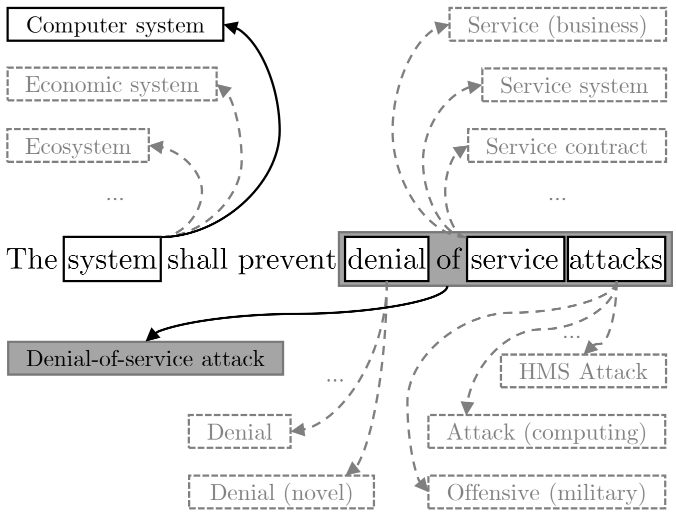

Fig. 1. Example of potential senses for the expressions in a requirement.

I. Introduction

Understanding requirements is essential for many software development tasks. Particularly in automatic processing of natural language requirements, the intents of textual expressions have to be understood and connected to some kind of knowledge representation. It is necessary to relate textual expressions to each other and to external knowledge to gain a deeper understanding of the requirements. For example, an automatic traceability link recovery approach could utilize the relations between expressions by exploiting super- and subconcepts. Expressions have different meanings depending on their context. Thus, they have to be disambiguated to prevent misinterpretations.

In requirements, many concepts are described by expressions that are composed of multiple words. However, most approaches only disambiguate single words instead of these multiword expressions (MWEs). These approaches fail at interpreting MWEs as a unit and miss connecting them to their correct senses. In the example "The system shall prevent denial of service attacks" in Figure 1 the information that the MWE *denial of service attack* is a network-based attack could be missed by associating the single words of the expression to their best fitting senses. Additionally, the example illustrates one of the difficulties of sense disambiguation as each word has many potential senses and it is hard to identify the correct sense in all contexts.

In computational linguistics, the task of determining the correct sense for each word is called word sense disambiguation (WSD). State-of-the-art approaches use supervised machine learning [1]–[5]. These approaches can only determine senses that were present during training. However, new domains and more domain-specific contexts are common in requirements engineering. Thus, costly labeling of training data and a retraining of model(s) is often needed for adaptation. An alternative to supervised approaches are knowledge-based approaches [6]–[10]. They can be adapted to new domains by providing an appropriate knowledge base and, thus, are more flexible and, consequently, more suitable for an application in requirements engineering.

We propose a two-tiered approach to overcome the issue of disambiguating multiword expressions in requirements documents. First, we use a conditional random field-based

(CRF) approach to detect MWEs in requirements documents. CRFs are undirected probabilistic models for segmenting and labeling sequences that offer various benefits such as relaxed assumptions compared to (hidden) Markov models (cf. [11]). We utilize the mwe-toolbox3 [12] to detect MWEs using CRFs. Then, we disambiguate the MWEs using a graph-based sense disambiguation approach. We use the knowledge-based WSD approach UKB [7] and extend it to utilize Wikipedia and to additionally detect partial senses.

We provide a dataset [13] including 997 requirement sentences annotated with multiword and single word expressions and their corresponding senses in Wikipedia and WordNet 3.1 [14]. For MWEs whose senses are not contained in the respective knowledge base the dataset additionally contains senses that fit the expression partly.

II. Related Work

The two research areas of particular relevance to our approach are the detection of MWEs and WSD.

The workshop series SemEval[1], that evolved from the Senseval series[2], has the goal to advance semantic analyses. The series targets the detection of MWEs in several tasks such as DiMSUM [15] and the PARSEME Verbal Multi-Word Expression Shared Task 2017 [16]. The DiMSUM task addresses the detection of minimal semantic units and their meanings [15]. Approaches need to combine labeling of MWEs and supersenses. Supersenses are generalized named entity classes for nouns (26 classes) and verbs (15 classes). In this task, approaches scored up to 57.7% F_1-score in a multi-domain evaluation. The approach UW-CSE by Hosseini et al. [17] achieves one of the best results using a conditional random field (CRF). The approach UFRGS&LIF by Cordeiro et al. [18] uses heuristic pattern-matching to detect MWEs. The results in both, MWE detection and supersense labeling, are good. Björne and Salakoski introduce UTU [19] that matches word sequences against given resources. On one hand, this approach is comparably weak at detecting MWEs. On the other hand, their classifier-based approach to choose supersenses performs on par with other approaches. The goal of the PARSEME Verbal Multi-Word Expression Shared Task 2017 [16] is to tackle verbal MWEs as they are rarely modelled due to their complexity. Noteworthy approaches are using neural networks (cf. [20]) or CRF sequence models (cf. [21]), whereby the CRF-based approach outperforms the neural network.

In the area of WSD, Arranz et al. study the impact of MWEs [22]. In contrast to traditional approaches that look for longest word-sequence matches, the authors use a knowledge-based approach to generate MWEs using WordNet. Using heuristics and incorporating lemmatization to increase performance, the approach shows promising performance on the Senseval-3 Task [23] with a precision of up to 81% and a recall of up to 84%.

[1]https://semeval.github.io/

[2]http://web.eecs.umich.edu/~mihalcea/senseval/

Non-MWE approaches for WSD can be separated into two main categories: knowledge-based WSD and supervised WSD. The former use knowledge bases to gain information about potential senses and select the closest candidate according to the given information. The latter are trained on data and learn common contexts of words etc. to disambiguate them. In general, supervised WSD approaches outperform knowledge-based ones, but need (usually expensive) training data and are only able to predict senses seen during training.

Knowledge-based approaches have in common that they often use a graph based on a knowledge base and calculate different metrics to select semantic interpretations. One example is UKB by Agirre et al. [6], [7]. The approach uses Personalized PageRank random walks over a semantic relations graph like WordNet. They achieve an F_1-score of up to 67.3% on the Senseval and SemEval tasks. Babelfy by Moro et al. [8] uses heuristics leveraging density within subgraphs. Some approaches, such as the approach by Chaplot and Salakhutdinov [9], use topic modelling to model topics within a document first and use this information in combination with knowledge bases to disambiguate word senses. The state-of-the-art for knowledge-based approaches by Wang et al. [10] integrates Latent Semantic Allocation (LSA). Their approach is tailored to WordNet, but uses Wikipedia to learn word representations.

For supervised WSD, approaches use different forms of machine learning. For example, SupWSD by Papandrea et al. [5] uses a support vector machine-based classifier on text features. Other approaches utilize glosses of WordNet to enhance the performance of their approach [2]–[4]. Recent approaches use BERT or similar transformer-based language models (cf. [2], [3]) and achieve state-of-the-art results. EWISER by Bevilacqua and Navigli [1] achieves an F_1-score of 80.1% on the Senseval and SemEval tasks.

III. Approach

In linguistics, MWEs can be categorized into seven subtypes: fixed expressions (such as *by and large*), (non-) decompositional idioms (e.g., *kick the bucket* or *let the cat out of the bag*), verb-particle constructions (such as *look up*), light verbs (e.g., *make a mistake*), proper names (such as *Karlsruhe Institute of Technology*), and compound nominals (e.g., *car park*) [24]. In the context of interpreting requirements, not all of these expression types are equally relevant. Idioms or fixed expressions are not concise and, thus, should not be used in requirements. Closely related are light verb constructs that should not be present in requirements as well, because they are highly idiosyncratic [24]. Moreover, verb-particle constructions are informative but do not yield information about the underlying concepts. Therefore, our approach disregards these subtypes. In contrast, compound nouns and proper names constitute a source of information about the concepts in the requirements and their interconnections. Consequently, they are highly informative. As a result, we focus on MWEs that are compound nouns and proper names.

TABLE I
OVERVIEW OF THE DATASET WITH INFORMATION ON THE NUMBER OF REQUIREMENT SENTENCES (REQSENTENCES), MWES, SINGLE WORD NOUNS AND PROPER NAMES WITH ANNOTATED SENSES (FURTHERSENSES) AND THEIR RESPECTIVE SHARES THAT HAVE A FITTING SENSE IN EITHER WIKIPEDIA (INWIKI) OR WORDNET (INWN).

				NFR															
Project	CM1	EBT	GANTT	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Total
ReqSentences	30	41	136	33	41	80	85	91	105	26	100	24	58	20	44	37	17	29	997
MWEs	50	45	161	28	42	211	194	110	121	41	115	54	33	19	50	64	22	48	1408
↪ inWiki	10	7	22	6	7	23	14	26	53	7	31	10	4	4	20	13	4	5	266
↪ inWN	2	2	2	3	4	14	5	7	25	3	14	1	4	1	6	6	1	2	102
furtherSenses	88	101	451	95	93	234	185	159	254	49	277	45	248	67	106	79	48	79	2658
↪ inWiki	43	83	391	78	79	179	130	128	225	35	252	35	211	50	93	63	41	54	2170
↪ inWN	35	88	396	94	91	219	176	150	235	45	247	42	248	63	90	66	38	69	2392

TABLE II
10-FOLD CROSS VALIDATION OF THE MWE DETECTION. CRF-PRT ALSO ALLOWS PARTIAL MATCHES.

Approach	Precision	Recall	F_1
CRF	0.914	0.856	0.884
CRF-PRT	0.949	0.881	0.913

In the following, we propose our two-tiered approach to disambiguate MWEs in requirements documents. We detect MWEs (along with their components) using a CRF-based tagger (cf. subsection III-A). Then, we disambiguate the expressions using a knowledge-based sense disambiguation that allows for partial sense disambiguation (cf. subsection III-B). This division allows us to choose different techniques (e.g., machine learning, knowledge-based) for detection and disambiguation.

A. Multiword Expression Detection

The first tier of our approach detects MWEs. We adapt CRF-based approaches that showed success on the SemEval tasks (cf. [15], [16]). Therefore, we train a single-chained CRF on a corpus of requirements using the mwe-toolkit3 [12]. We use the default feature set of the mwe-toolkit3 and $c2 = 1$ for L2 regularization. L2 regularization is a technique to prevent overfitting by regularizing parameters. We gathered a dataset consisting of 774 requirements from the CM1, EBT, and GANTT datasets retrieved from the Center of Excellence for Software & Systems Traceability (CoEST)[3] and the NFR dataset [25] with 18 projects in total (c.f. Table I). We tag MWEs manually in the DiMSUM format [15], resulting in 1408 MWEs. During preprocessing, we also annotate part of speech and lemma to each word. We make the dataset publicly available [13].

1) Evaluation: We evaluate the MWE detection with a random 10-fold cross validation on the dataset presented in Table I. Therefore, we test the tagger ten times on a tenth of the requirement sentences in the dataset and train on the remainder. There are two kinds of results in Table II: results that consider only perfect matches as true positive (CRF) and results that also reward partial matches (CRF +PRT). A partial match exists if preceding components of a MWE are omitted, e.g., the model only detects *player statistics* instead of *NHL player statistics*.

The detection achieves a high precision of 91.4% while still providing a good recall of 85.6%. This results in a promising F_1-score of 88.4%. In comparison, the CRF-based approach of Hosseini et al. [17] achieved an F_1-score of 61.1% on the SemEval-2016 task. This is likely caused by the fact that requirements consist of rather short and precise sentences compared to the sentences in the SemEval-2016 corpus. The errors of the approach can often be attributed to rare proper names with special symbols such as *LAST_BOOT_IVEC location*. Another source of errors are MWEs that consist of adjectives and nouns (e.g., *offensive player*). During training, this type of MWE is rarely seen, as most MWEs are combinations of nouns and proper names. Therefore, better training data could decrease this type of error.

If we reward partial matches as well, precision and recall increase both by 3%. In many cases, partially detected MWEs still provide helpful information for WSD of the entity, e.g., detecting *player statistics* instead of *NHL player statistics*. The result of an F_1-score of 91.3% shows that our MWE detection is a promising building block for sense disambiguation.

B. Knowledge-based Sense Disambiguation

Requirements usually contain many domain-specific expressions. Therefore, WSD approaches for requirements need to adapt to different domains. Supervised approaches achieve best results on benchmarks, but can not be easily adapted to new or specific knowledge/domains. They need to be (re-)trained on annotated, domain-specific training data.

Knowledge-based approaches can be adapted to different domains by exchanging the underlying knowledge base. Given that advantage, graph-based approaches present a particularly flexible solution as knowledge is often depicted with concepts and relations between concepts. Therefore, we propose a graph-based approach for WSD in requirements, using a knowledge base that fits best for a given domain. Alternatively, a combination of, e.g., domain knowledge and general knowledge by merging their respective graphs is possible.

[3]http://coest.org/

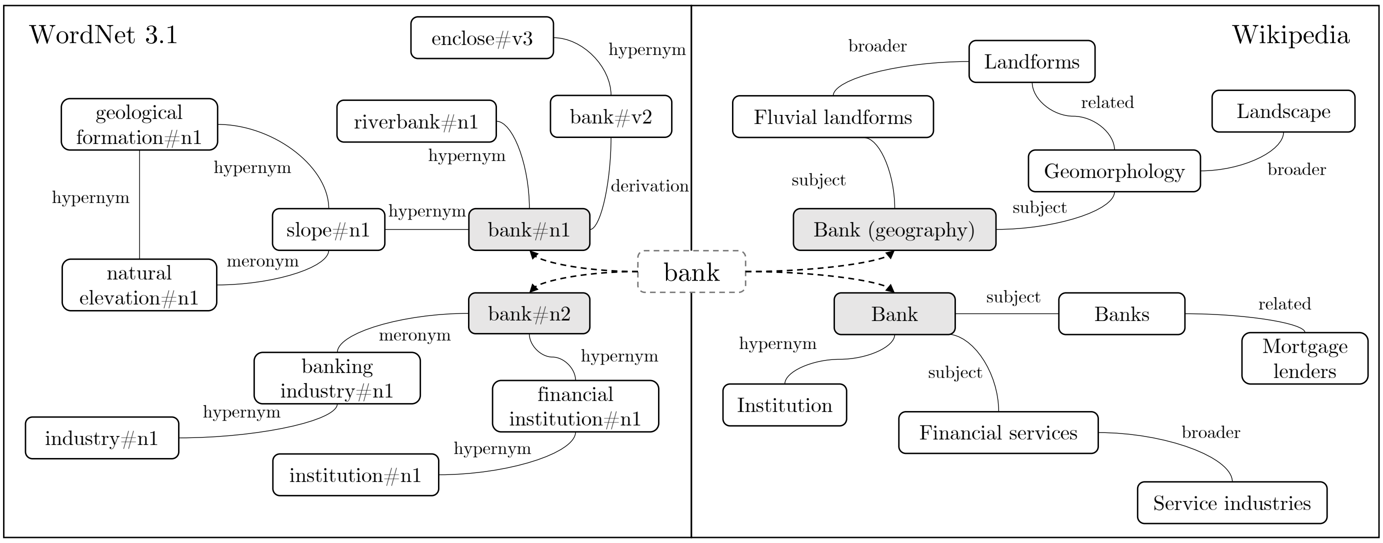

Fig. 2. Excerpt of WordNet and Wikipedia semantic graphs for different senses of *bank*.

We employ the graph-based sense disambiguation tool UKB [6], [7]. It uses Personalized PageRank random walks on semantic relation graphs. The Personalized PageRank weights nodes that are connected to context words higher and, thus, incorporates context information.

UKB requires a semantic graph and a dictionary as input. The dictionary maps lemmas of expressions to their possible senses. The semantic graph contains concepts and their semantic relations. Originally, UKB uses a WordNet 3.0 graph with hypernyms, meronyms, antonyms, derivations, and senses of the words in the glosses as semantic relations. We adapted the approach by Agirre et al. to provide a graph based on WordNet 3.1 instead of WordNet 3.0 as WordNet 3.1 covers more senses. However, we had to omit the extended gloss relations as they are not available for WordNet 3.1. The left side of Figure 2 shows an excerpt of the WordNet 3.1 semantic graph for different senses of *bank*. In this example, *bank* in a geographical sense is connected to *slope* via hypernymy. A *slope* itself is a part of a natural elevation that is a geological formation. All of these senses again are connected to further senses; the graph spans the entire knowledge base.

However, we believe WordNet is not the most fitting general knowledge base for requirements. It is mainly focused on common expressions in news texts and literature. We believe that Wikipedia is more suitable because it contains more concepts that are relevant for requirements. Therefore, we extracted a semantic relation graph from Wikipedia using DBpedia resources [26] [4]. We use the relations in the Linked Hypernym Dataset [27] and the category information of Wikipedia articles as relations to replicate the semantic relations in WordNet. The former contains hypernym relations, e.g., a *bank* is an *institution*. The latter contains the relations `subject`, `broader` and `related`. `subject` relates an article to its category. `broader` and `related` identify hypernym and other relations between categories. The right side of Figure 2 shows an excerpt of the resulting semantic graph for two senses of the word *bank*. The senses are represented by articles in Wikipedia. The article on *banks* in a geographical sense has *geomorphology* as a category which is related to the category *landforms* and additionally has *landscape* as a hypernym. We construct two versions of the Wikipedia graph: one only uses the Linked Hypernym Dataset and the other additionally contains category relations. We leverage the links in Wikipedia disambiguation pages to provide a dictionary that maps expressions to their possible senses. Additionally, UKB is able to use sense frequencies as edge weights. We use the number of articles that include a link to the given article as frequency for Wikipedia senses.

UKB provides two versions of the Personalized PageRank: `ppr_w2w` applies PageRank to each word whereas `ppr` applies PageRank to each sentence only once. We use the recommended number of iterations (30) and damping factor (0.85). As context for each disambiguation step, we only use the enclosing sentence that the MWE is contained in. Experiments with a broader context of two surrounding sentences resulted in a small improvement for Wikipedia but a worse performance on WordNet.

For disambiguation with WordNet, we use two versions that each consider different contexts. The first only includes nouns and MWEs as context expressions. This variant is comparable to the Wikipedia version, as Wikipedia only includes senses for concepts that consist of nouns or MWEs. The second variant includes all nouns, verbs, adjectives, adverbs, and MWEs in the sentence. This aligns with WordNet that includes senses for these word types.

[4]https://databus.dbpedia.org/dbpedia/transition/linked-hypernyms/2019.02.10, https://databus.dbpedia.org/dbpedia/generic/categories/2019.08.30, https://databus.dbpedia.org/dbpedia/generic/disambiguations/2019.08.30, and https://databus.dbpedia.org/dbpedia/generic/wikipedia-links/2019.08.30

TABLE III
RESULTS OF WIKIPEDIA-BASED VARIANTS IF MISSING SENSES IN THE KNOWLEDGE BASE ARE COUNTED AS FALSE NEGATIVES. MFS DESCRIBES THE MOST FREQUENT SENSE BASELINE.

Graph	Hypernyms + Categories				Hypernyms				
PageRank	ppr		ppr_w2w		ppr		ppr_w2w		
Frequencies	✓	×	✓	×	✓	×	✓	×	MFS
Precision	0.227	0.295	**0.405**	0.378	0.360	0.385	0.319	0.276	0.218
Recall	0.213	0.276	**0.380**	0.355	0.330	0.352	0.291	0.253	0.204
F_1	0.220	0.285	**0.392**	0.366	0.344	0.368	0.304	0.264	0.211

TABLE IV
RESULTS OF WORDNET-BASED VARIANTS IF MISSING SENSES IN THE KNOWLEDGE BASE ARE COUNTED AS FALSE NEGATIVES. CONTEXTS CONSIST OF NOUNS (N), MULTIWORD EXPRESSIONS (MWE), VERBS (V), ADJECTIVES (JJ) OR ADVERBS (RB).

Context	N, V, JJ, RB, MWE				N, MWE				
PageRank	ppr		ppr_w2w		ppr		ppr_w2w		
Frequencies	✓	×	✓	×	✓	×	✓	×	MFS
Precision	**0.403**	0.269	0.374	0.310	0.399	0.267	0.368	0.305	0.340
Recall	**0.364**	0.243	0.339	0.280	0.361	0.241	0.333	0.276	0.308
F_1	**0.383**	0.255	0.356	0.294	0.379	0.253	0.350	0.290	0.324

Besides MWEs, we also disambiguate nouns and proper names that consist of only one word. We need their correct senses to provide a context for the MWE sense disambiguation. Furthermore, we need their senses anyways for automatic requirement interpretation.

We expect that not all MWEs are covered entirely by the used knowledge bases. We cover this with a simple heuristic to disambiguate MWEs partially. If UKB cannot find a sense for the entire MWE, we iteratively reduce the expression by one word from the left. Thus, if no sense for *NHL player statistics* is contained in the knowledge base, we try to disambiguate *player statistics* and, finally, *statistics*.

1) Evaluation: For evaluation, we extended the dataset in Table I with sense information from Wikipedia and WordNet 3.1. For each MWE and single word noun or proper name, we annotate the most specific sense in the knowledge bases. If the correct sense of a MWE is not contained in the knowledge base, we search for partial senses that are most fitting to the complete MWE, annotate these senses instead and mark them as partial. If still no correct sense is available, we annotate this as a deficit of the knowledge base. Table I gives an overview of the total amount of tagged expressions and their coverage in Wikipedia and WordNet. Only 102 out of 1408 MWEs have a correct sense in WordNet, 266 in Wikipedia. However, 1072 and 778 MWEs have partial senses respectively. Thus, we can confirm our hypothesis that Wikipedia contains more complete MWE senses. The coverage is more promising for single word nouns and proper names: WordNet contains 2392, Wikipedia 2170 correct senses out of 2658 expressions. Usually, single word nouns are less domain-specific and, thus, more likely included in a general knowledge base.

First, we want to determine the best configuration of our approach for each knowledge base. Table III shows the results obtained on all annotated senses with the two Wikipedia-based graphs. Missing senses in the knowledge base and partial senses are counted as false negatives. We get the best performance with the hypernyms and categories graph using `ppr_w2w` and sense frequencies. This configuration outperforms the F_1-score of the most frequent sense baseline (MFS) by over 18 percentage points. However, in comparison to the best configuration on the hypernyms-only graph (ppr without sense frequencies), the improvement is only 0.024.

TABLE V
RESULTS OF THE BEST PERFORMING CONFIGURATIONS IF MISSING SENSES IN THE KNOWLEDGE BASE ARE NOT COUNTED AS FALSE NEGATIVES (*-KB*) AND/OR PARTIAL SENSES FOR THE MWES ARE COUNTED AS CORRECT (*-PS*). MFS *-KB* DESCRIBES THE MOST FREQUENT SENSE BASELINE IN *-KB* SETTING. FOR BABELFY ADDITIONAL SENSES WERE NOT COUNTED AS FALSE POSITIVES.

Graph	Method	Precision	Recall	F_1
Wikipedia	ppr_w2w + frequencies	0.405	0.380	0.392
	-KB	0.405	0.635	0.495
	-PS	0.514	0.482	0.497
	-KB -PS	0.514	0.610	0.558
	Babelfy *-KB -PS*	0.532	0.368	0.435
	MFS *-KB*	0.218	0.341	0.266
WordNet	ppr + frequencies	0.403	0.364	0.383
	-KB	0.403	0.594	0.480
	-PS	0.557	0.504	0.529
	-KB -PS	0.557	0.574	0.566
	MFS *-KB*	0.340	0.502	0.406

For WordNet, the results in Table IV indicate that the best

TABLE VI
MWE-ONLY RESULTS IF MISSING SENSES IN THE KNOWLEDGE BASE ARE NOT COUNTED AS FALSE NEGATIVES (*-KB*) AND/OR PARTIAL SENSES FOR THE MWES ARE COUNTED AS CORRECT (*-PS*). FOR BABELFY ADDITIONAL SENSES WERE NOT COUNTED AS FALSE POSITIVES.

Graph	Method	Precision	Recall	F_1
Wikipedia	ppr_w2w + frequencies	0.159	0.145	0.152
	-KB	0.159	0.767	0.263
	-PS	0.481	0.439	0.459
	-KB -PS	0.481	0.592	0.531
	Babelfy	0.125	0.082	0.099
	Babelfy *-KB -PS*	0.274	0.245	0.258
WordNet	ppr + frequencies	0.062	0.056	0.059
	-KB	0.062	0.775	0.114
	-PS	0.503	0.458	0.480
	-KB -PS	0.503	0.549	0.525

configuration is `ppr` with sense frequencies and using all context senses. The configuration that can be compared to Wikipedia (Nouns and MWE) performs only slightly worse. However, the results on WordNet are close to the MFS baseline (six percentage points difference for F_1-score). For WordNet, the `ppr` setting outperforms the `ppr_w2w`. This contradicts the statement of Agirre et al., that `ppr_w2w` is slower but more precise (cf. [6]). We attribute this effect to the aforementioned reduced number of relations in our WordNet 3.1 graph. Our WordNet graph has only one third the edges of the graph used by Agirre et al. The same effect can be noted for the hypernyms-only graph for Wikipedia that has a lower number of relations as well.

The rather low performance on both knowledge bases can partly be explained by the coverage of correct senses in the knowledge base. The ceiling for recall is 59.9% on Wikipedia and 61.3% on WordNet. As a result of missing senses, many partial senses are annotated, which reduces the precision in this evaluation setting. Therefore, we present the results of the best performing configurations if deficits of the knowledge base are not counted (*-KB*) and/or partial senses for the MWEs are counted as a hit (*-PS*) in Table V. The recall increases to over 59% on both knowledge bases. The acceptance of partial senses additionally increases the precision to over 51%. This result indicates that our approach is able to detect partially correct senses for cases where the knowledge base misses an entirely fitting sense. On the *-KB* setting, our approach again outperforms the most frequent sense baselines. On Wikipedia, we can additionally compare our approach to Babelfy [8]. In the *-KB -PS* setting our approach outperforms Babelfy by over ten percentage points in F_1-score and recall. Note that Babelfy tends to annotate multiple senses to choose from. We do not count these additional senses as false positives. Otherwise, the performance of Babelfy would be even lower.

As our approach aims at disambiguating MWEs, we present in Table VI the results of our approach and Babelfy solely on the MWEs of the dataset. The results are very low in the standard evaluation setting. Again, recall is lowered by the low

TABLE VII
COMBINED RESULTS IF MISSING SENSES IN THE KNOWLEDGE BASE ARE NOT COUNTED AS FALSE NEGATIVES (*-KB*) AND/OR PARTIAL SENSES FOR THE MWES ARE COUNTED AS CORRECT (*-PS*).

Graph	Method	Precision	Recall	F_1
Wikipedia	ppr_w2w + frequencies	0.390	0.363	0.376
	-KB	0.390	0.606	0.474
	-PS	0.485	0.452	0.468
	-KB -PS	0.485	0.571	0.525
WordNet	ppr + frequencies	0.394	0.352	0.372
	-KB	0.394	0.574	0.468
	-PS	0.533	0.476	0.503
	-KB -PS	0.533	0.542	0.537

amount of covered senses in the knowledge base. Looking at the results that acknowledge partial senses, we conclude that our approach is capable of detecting fitting senses. The respective F_1-scores improve by up to 38 and 47 percentage points on Wikipedia and WordNet. In the *-KB -PS* setting the results are even comparable to the overall results. Moreover, our approach clearly outperforms Babelfy in both settings. This is probably because Babelfy does not focus on MWEs and weights WSD of single words higher.

To answer the question of how well our two-tiered approach performs, we perform an experiment that combines both steps: We employ our MWE detection instead of using the MWEs from our gold standard and additionally disambiguate all remaining single word nouns. We take the detection results from the same ten folds we used in subsection III-A. Table VII shows the results. The results decrease slightly, as expected of an approach building upon a detection performance of 88% F_1-score. As no major decline is present, we conclude that our two-tiered approach is promising for disambiguating MWEs.

IV. THREATS TO VALIDITY

There are some potential threats to validity of our research and experimental design that we discuss in the following.

a) External Validity: The probably most major threat to validity of our work concerns external validity. The chosen dataset for evaluation might not be representative for requirements in general. It covers 18 different projects that mostly stem from academic projects. However, the projects are widely used and accepted in the research community, are of different sizes, and cover different domains.

b) Internal Validity: Another threat might be the fact that the gold standard was created with the approach in mind. It thus may suffer from experimenter bias. Additionally, determining the correct senses for natural language expressions can be a challenging task for humans as well, thus the dataset may include errors. We try to mitigate this risk by publishing the dataset [13], so that everyone can reproduce our results and findings.

V. CONCLUSION

In this paper, we presented a knowledge-based approach to disambiguate multiword expressions (MWEs) in requirements documents. The combination of a machine learning-based approach for multiword expression detection together with a graph-based approach for sense disambiguation achieves a high detection rate. Additionally, the approach might be adapted to new domains and knowledge by exchanging the semantic relation graph. However, in this paper we focused on the two general knowledge bases Wikipedia and WordNet.

We evaluated both parts of our approach individually and in combination. Our evaluation of the multiword expression detection shows that the approach achieves high accuracy when trained on requirements from different domains. The evaluation of the sense disambiguation was performed with Wikipedia and WordNet 3.1 as knowledge bases. Our approach outperforms other knowledge-based approaches on requirements. As an advantage, the approach also detects partial senses for MWEs. Combined, the performance of our approach does not decrease much, showing that the overall approach is promising for the disambiguation of MWEs.

However, the evaluation showed that many senses of MWEs are not covered by the selected knowledge bases. Enhancing these existing general knowledge bases or providing domain specific knowledge bases requires a significant amount of effort. Thus, the quality and availability of fitting knowledge bases clearly limits the applicability of our approach.

In future extensions, we plan to integrate further knowledge bases and experiment with combining domain and general knowledge bases for better coverage. Moreover, detection approaches based on language models and transfer learning such as BERT might yield even higher accuracies.

REFERENCES

[1] M. Bevilacqua and R. Navigli, "Breaking Through the 80% Glass Ceiling: Raising the State of the Art in Word Sense Disambiguation by Incorporating Knowledge Graph Information," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp. 2854–2864.

[2] T. Blevins and L. Zettlemoyer, "Moving Down the Long Tail of Word Sense Disambiguation with Gloss Informed Bi-encoders," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Jul. 2020, pp. 1006–1017.

[3] L. Huang, C. Sun, X. Qiu, and X. Huang, "GlossBERT: BERT for Word Sense Disambiguation with Gloss Knowledge," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3509–3514.

[4] F. Luo, T. Liu, Q. Xia, B. Chang, and Z. Sui, "Incorporating Glosses into Neural Word Sense Disambiguation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 2473–2482.

[5] S. Papandrea, A. Raganato, and C. Delli Bovi, "SupWSD: A Flexible Toolkit for Supervised Word Sense Disambiguation," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 103–108.

[6] E. Agirre, O. López de Lacalle, and A. Soroa, "Random Walks for Knowledge-Based Word Sense Disambiguation," *Computational Linguistics*, vol. 40, no. 1, pp. 57–84, Mar. 2014.

[7] ——, "The risk of sub-optimal use of Open Source NLP Software: UKB is inadvertently state-of-the-art in knowledge-based WSD," in *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Melbourne, Australia: Association for Computational Linguistics, Jul. 2018, pp. 29–33.

[8] A. Moro, A. Raganato, and R. Navigli, "Entity Linking meets Word Sense Disambiguation: A Unified Approach," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 231–244, 2014.

[9] D. S. Chaplot and R. Salakhutdinov, "Knowledge-based Word Sense Disambiguation using Topic Models," in *Thirty-Second AAAI Conference on Artificial Intelligence*, Apr. 2018.

[10] Y. Wang, M. Wang, and H. Fujita, "Word Sense Disambiguation: A comprehensive knowledge exploitation framework," *Knowledge-Based Systems*, vol. 190, p. 105030, Feb. 2020.

[11] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[12] C. Ramisch, *Multiword Expressions Acquisition: A Generic and Open Framework*, ser. Theory and Applications of Natural Language Processing. Springer International Publishing, 2015.

[13] T. Hey, J. Keim, and W. F. Tichy, "Dataset of "Knowledge-based Sense Disambiguation of Multiword Expressions in Requirements Documents"," Aug. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5167247

[14] C. Fellbaum, *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[15] N. Schneider, D. Hovy, A. Johannsen, and M. Carpuat, "SemEval-2016 Task 10: Detecting Minimal Semantic Units and their Meanings (DiMSUM)," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 546–559.

[16] A. Savary, C. Ramisch, S. R. Cordeiro, F. Sangati, V. Vincze, B. Qasemi Zadeh, M. Candito, F. Cap, V. Giouli, I. Stoyanova *et al.*, "The parseme shared task on automatic identification of verbal multiword expressions," in *The 13th Workshop on Multiword Expression at EACL*, 2017, pp. 31–47.

[17] M. J. Hosseini, N. A. Smith, and S.-I. Lee, "UW-CSE at SemEval-2016 Task 10: Detecting Multiword Expressions and Supersenses using Double-Chained Conditional Random Fields," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California, 2016, pp. 931–936.

[18] S. Cordeiro, C. Ramisch, and A. Villavicencio, "UFRGS&LIF at SemEval-2016 Task 10: Rule-Based MWE Identification and Predominant-Supersense Tagging," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, San Diego, California, Jun. 2016, pp. 910–917.

[19] J. Björne and T. Salakoski, "UTU at SemEval-2016 Task 10: Binary Classification for Expression Detection (BCED)," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 925–930.

[20] N. Klyueva, A. Doucet, and M. Straka, "Neural Networks for Multi-Word Expression Detection," in *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 60–65.

[21] A. Maldonado, L. Han, E. Moreau, A. Alsulaimani, K. Chowdhury, C. Vogel, and Q. Liu, "Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking," 2017.

[22] V. Arranz, J. Atserias, and M. Castillo, "Multiwords and Word Sense Disambiguation," in *Computational Linguistics and Intelligent Text Processing*, ser. Lecture Notes in Computer Science, A. Gelbukh, Ed. Berlin, Heidelberg: Springer, 2005, pp. 250–262.

[23] K. Litkowski, "Senseval-3 task: Word Sense Disambiguation of WordNet glosses," in *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain, Jul. 2004, pp. 13–16.

[24] I. A. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger, "Multiword Expressions: A Pain in the Neck for NLP," in *Computational Linguistics and Intelligent Text Processing*, ser. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2002, pp. 1–15.

[25] Jane Cleland-Huang, S. Mazrouee, H. Liguo, and D. Port, "Nfr," Mar. 2007.

[26] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer, "DBpedia – A large-scale, multilingual knowledge base extracted from Wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.

[27] T. Kliegr, "Linked Hypernyms: Enriching DBpedia with Targeted Hypernym Discovery," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3199181, 2015.

Karlsruher Institut für Technologie

OPEN ACCESS

Repository KITopen

Dies ist ein Postprint/begutachtetes Manuskript.

Empfohlene Zitierung:

Hey, T.; Keim, J.; Tichy, W. F.
Knowledge-based Sense Disambiguation of Multiword Expressions in Requirements Documents.
2021. 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW), Notre Dame, IN, USA, 20-24 Sept. 2021
doi: 10.5445/IR/1000139591

Zitierung der Originalveröffentlichung:

Hey, T.; Keim, J.; Tichy, W. F.
Knowledge-based Sense Disambiguation of Multiword Expressions in Requirements Documents.
2021. 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW), Notre Dame, IN, USA, 20-24 Sept. 2021, 70–76, Institute of Electrical and Electronics Engineers (IEEE).
doi:10.1109/REW53955.2021.00017

Lizenzinformationen: KITopen-Lizenz