Definition Frames: Using Definitions for Hybrid Concept Representations

Evangelia Spiliopoulou **Artidoro Pagnoni** **Eduard Hovy**
Language Technologies Institute
Carnegie Mellon University
{espiliop,apagnoni,hovy}@cs.cmu.edu

Abstract

Advances in word representations have shown tremendous improvements in downstream NLP tasks, but lack semantic interpretability. In this paper, we introduce Definition Frames (DF), a matrix distributed representation extracted from definitions, where each dimension is semantically interpretable. DF dimensions correspond to the Qualia structure relations (Boguraev and Pustejovsky, 1990): a set of relations that uniquely define a term. Our results show that DFs have competitive performance with other distributional semantic approaches on word similarity tasks.

1 Introduction

Ontologies have been widely used in lexical semantics to organize and represent knowledge. Carefully built by experts, they contain semantically meaningful information in the form of relations between concepts. However, being manually constructed, they struggle to assimilate new information.

Compared to ontologies, distributed representations are fully automated and can be fine-tuned for new tasks. Despite their exceptional performance, most distributional methods do not have an explicit semantic interpretation. The resulting representations encode a tremendous amount of information, but afford no way to interpret what this information is and how it relates to the concept. Thus, one cannot choose which type of information is useful for a specific task, unless one has a lot of data and resources to fine-tune. Although a few approaches have tried to bridge the gap between semantics and distributed representations (Faruqui et al., 2015; Mrkšić et al., 2017), (1) they only encode information from ontologies, which are not extensible, and (2) the final representations are still not semantically meaningful.

Motivated by these problems, we introduce a novel hybrid representation called **Definition Frames** (DF), which encode semantic information extracted from definitions. DFs are matrix representations, where each row corresponds to a particular relation. The set of the relations used is based on the Qualia structure suggested in Boguraev and Postojovsky (1990), and they are extracted automatically from definitions via a domain-adaptation approach. To the best of our knowledge, DF is the first hybrid representation, combining an explicit structure through semantically meaningful rows, while still being decomposed into distributional vectors.

2 Prior Work

Prior research on lexical semantics has established a set of relations that are sufficient to uniquely define a concept. Such work includes the Qualia structure (Boguraev and Pustejovsky, 1990) and the generative lexicon theory (Pustejovsky, 1991). Other related work includes ontological approaches (Baker et al., 1998; Miller, 1995; Lenat, 1995; Speer and Havasi, 2012) and more fine-grained definition-based frames like Semagrams (Moerdijk and others, 2008).

In distributional semantics, approaches including GloVe (Pennington et al., 2014), word2vec (Mikolov et al., 2013), and fastText (Bojanowski et al., 2017) obtain generic word embeddings by pre-training on large corpora. Recent work focused on context-sensitive embeddings like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), which achieve significant improvements in downstream NLP tasks.

Proceedings of the 28th International Conference on Computational Linguistics, pages 3060–3068
Barcelona, Spain (Online), December 8-13, 2020

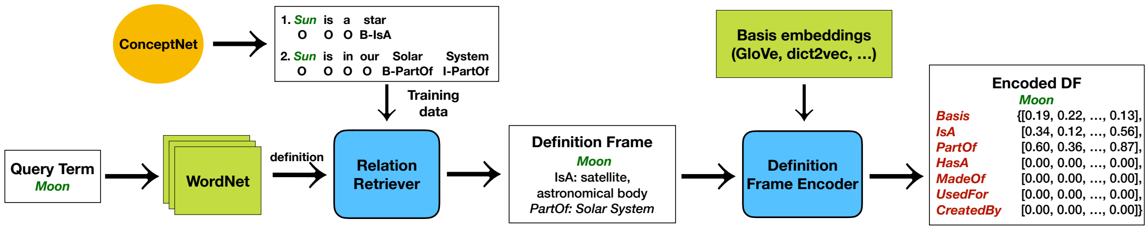

Figure 1: Architecture diagram.

Earlier work on definitions extracted the type of a concept (*Genus*) and the relations distinguishing it from other members of the same type (*Differentia*) via syntax and string matching heuristics (Binot and Jensen, 1993; Calzolari, 1984; Chodorow et al., 1985). Recent approaches directly encoded definitions to distributed representations. Tissier (2017) obtained embeddings via a skip-gram model trained on definitions, while Bosc (2018) used an auto-encoder. Other work includes definition generation (Noraset et al., 2017), binary classification of sentences on whether they are definitional (Anke and Schockaert, 2018), reverse dictionary look-up (Hill et al., 2016; Zock and Bilac, 2004), and extraction of hypernymy relations from definitions using syntactic patterns (Boella and Di Caro, 2013).

3 Approach

Our framework consists of two parts: the Relation Retriever and the Definition Frame (DF) Encoder. The WordNet definition for any given term is used by the Relation Retriever model to extract the Qualia structure relations. The set of extracted terms pertaining to these relations form the Definition Frame. The DF Encoder encodes this output to a distributed matrix representation, which can be used in downstream NLP tasks.

Qualia Structure The Qualia structure (formal, constitutive, telic, and origin) is defined as the complete modes of explanation associated with an entity (Boguraev and Pustejovsky, 1990; Pustejovsky, 1991). These relations suffice to uniquely and completely define a concept. In fact, several Relation Extraction tasks (Hendrickx et al., 2009; Gábor et al., 2018) contain relations similar to Qualia describing the type (*isA*), structure (*madeOf, partOf, hasA*), function (*usedFor*), or provenance (*createdBy*) of a concept.

Qualia	Relation	# Wikipedia Def.	# WordNet Def.	WordNet Overlap
Formal	IsA	235	146	59% (87/146)
Constitutive / Structure	PartOf	82	57	2% (1/57)
	HasA	39	33	6% (2/33)
	MadeOf	27	19	5% (1/19)
Telic / Function	UsedFor	59	54	0% (0/54)
Origin / Provenance	CreatedBy	26	17	0% (0/17)

Table 1: Annotated Relations for 300 Wikipedia and 150 WordNet definitions. *WordNet Overlap* indicates the number of relations expressed in the definition that were present in the WordNet ontology.

To automatically extract the Qualia structure of a term, we use dictionary definitions, as they uniquely describe a term. We confirm the prevalence of those relations in definitions by annotating 300 Wikipedia and 150 WordNet definitions, chosen at random from nominal terms in WordNet (Table 1). We empirically find that WordNet definitions express more relations than the hypernymy (*isA*) and meronymy (*madeOf, partOf, hasA*) relations directly encoded in the WordNet ontology (usedFor and createdBy relations are not part of WordNet ontology). Furthermore, as shown in Table 1, we observe that meronymy relations are more prevalent in WordNet definitions compared to the ontology.

Training Data Because there are no definitions annotated with Qualia structure and Relation Extraction datasets (Hendrickx et al., 2009; Gábor et al., 2018) are very domain specific without encoding general knowledge, we deploy a domain adaptation technique. We use ConceptNet to pre-train the Relation Retriever model (section 3) and then fine-tune it on and apply it to WordNet definitions. We fine-tune on a set of 150 manual annotations, since WordNet definitions tend to have more complex sentences than the ones in ConceptNet.

ConceptNet (Speer and Havasi, 2012) is a general purpose ontology that contains relations between pairs of concepts, accompanied by a small source-sentence. Figure 1 shows that the Concept-query *Sun* is linked to two sentences (*Sun is a star* and *Sun is in our solar system*) from ConceptNet with the corresponding relations *isA* and *partOf*. The training data for the Relation Retriever is composed of all ConceptNet source-sentences that contain one of the Qualia structure relations.

Extracting Definition Frames [1] The Relation Retriever uses the WordNet definition of a term to extract words that are related to that term via a Qualia-type relation. The set of extracted relations with their corresponding related words form the **Definition Frame** (DF). More specifically, we define a Definition Frame for a term t as $F_t = \{r_1 : S_1, r_2 : S_2, .., r_k : S_k\}$, where $r_i \in \{$ *isA, usedFor, partOf, hasA, madeOf, createdBy* $\}$ and S_i is the set of words related to t via the relation r_i. For example, to extract the DF for *moon* (Figure 1), we use the WordNet definition of *moon* as input. The Relation Retriever extracts the terms that are related to *moon* via a Qualia-structure relation (i.e. *satellite*, *astronomical body* and *solar system*). These terms with their corresponding relations constitute the Definition Frame F_{moon}. More examples of Definition Frames are shown in Table 2.

Word 1	**Definition Frame, word 1**	**Word 2**	**Definition Frame word 2**	**Relatedness**
shore	IsA: land, edge PartOf: body, water	sea	IsA: body PartOf: ocean, salt, water CreatedBy: land	0.86
wool	IsA: fabric MadeOf: hair, sheep	fabric	IsA: artifact MadeOf: weaving HasA: fibers CreatedBy: felting, knitting	0.86
restaurant	IsA: building, people UsedFor: eat	dinner	IsA: main, meal PartOf: day, evening, midday	0.86
day	IsA: time UsedFor: earth, make, complete, rotation	dusk	IsA: time PartOf: day, following, sunset	0.76
dress	IsA: one-piece, garment UsedFor: woman HasA: skirt, bodice	bride	IsA: woman CreatedBy: married	0.76
feather	IsA: light, horny, waterproof, structure PartOf: external, covering	hawk	IsA: diurnal, bird HasA: short, rounded, wings	0.82
orange	IsA: round, yellow, orange, fruit PartOf: citrus, trees	fruit	IsA: ripened, reproductive, body PartOf: seed, plant	0.82
harbour	IsA: sheltered, port, ships UsedFor: discharge, cargo	boat	IsA: small, vessel UsedFor: travel, water	0.76

Table 2: Extracted Definition Frames (before encoding) for pairs with high Relatedness score (MEN dataset). The Relatedness score, is the ground truth score, as noted in the original dataset. We observe that the two terms share characteristics of their Definition Frame, like being part of each other's frame or having common related terms.

The Relation Retriever uses a BiLSTM model to extract the relations from each sentence. The task is formulated as a sequence tagging problem where we identify both the relation type and the related entities, and optimizes the cross-entropy loss. For model selection, we perform experiments with strong baseline architectures for RE tasks (BiLSTM, BERT-BiLSTM, BiLSTM-CNN). The Relation Retriever obtains F1 = 0.97 on ConceptNet test data (Appendix A.1).

[1] Code available in github.com/spilioeve/Definition-Frames.

The Definition Frame is encoded via the DF Encoder into a matrix where each row w_i corresponds to one of the Qualia relations. The DF Encoder uses an embedding space ($Basis$) to construct each row vector w_i. Note that $Basis$ can be any distributional embedding model. Given a DF F_t, we define w_i as the average of word embeddings from the set of related terms S_i through relation r_i:

$$w_i = \frac{1}{|S_i|} \sum_{s \in S_i} Basis(s)$$

where $Basis(s)$ is the embedding for word s. We include an additional row for the $Basis$ vector of the term itself. This encoding of DF maintains a semantically meaningful structure as each row always corresponds to the same relation. If no terms are extracted for a relation, we use the zero vector of appropriate size. An example of the encoded DF$_{moon}$ is shown in Figure 1, where each dimension corresponds to a unique relation like *isA* and *partOf*.

4 Experiments

Word-Similarity Task We perform experiments on benchmark word-similarity datasets provided by Faruqui (2014): SimLex999 (Hill et al., 2015), MC30 (Miller and Charles, 1991), RG65 (Rubenstein and Goodenough, 1965), WS353 (Finkelstein et al., 2002) and MEN (Bruni et al., 2012). Following Agirre (2009), we split them into word-similarity (WS-Sim, SimLex999, MC30, RG65) and word-relatedness (WS-Rel, MEN) datasets, as they evaluate different semantic affinities. We only consider nominal terms that exist in WordNet and report Spearman's correlation ρ. We perform experiments with three types of embeddings used as $Basis$: GloVe (Pennington et al., 2014), dict2vec trained on Wikipedia (Tissier et al., 2017), and retrofit embeddings (Faruqui et al., 2015) based on GloVe. Since the task comprises of pairs of words without any context, we do not compare against context-based representations.

Ablation Study We perform an ablation study by varying the set of relations used in DF. In this study, both $Basis$ and DF are encoded with dict2vec, as it achieves the best performance (Table 3). The goal of this study is to measure how each extracted relation affects the performance of DF in word similarity tasks. The results (details in Appendix A.2) show that, for similarity tasks, pruning relations sometimes improves performance over both the original DF (with all relations) and the $Basis$ embeddings. However, we observe that DFs consistently have worse performance than $Basis$ in relatedness tasks, particularly in the MEN dataset. As we further discuss in detail in Section 4, although DFs capture relatedness, this is not reflected when using the cosine similarity metric directly, since it cannot compare information across different dimensions. For example, consider the pair (*car*, *wheel*). If we compare row-vectors of DF_{wheel} and DF_{car} for each relation separately, the representations are very different. Each Qualia structure relation defining *car* and *wheel* is different for the two terms. However, the Structure dimension of DF_{car} would contain the information that *wheel* is part (meronym) of *car*, thus it should be compared to the $Basis$ dimension of DF_{wheel}.

Datasets	**GloVe**				**Dict2vec**				**Retrofit**			
	Basis	Basis*	**DF**	**DF***	Basis	Basis*	**DF**	**DF***	Basis	Basis*	**DF**	**DF***
Similarity CV	0.39	0.50	0.35	**0.53**	0.53	0.52	0.45	**0.56**	0.44	**0.59**	0.35	0.56
Relatedness CV	0.68	0.77	0.38	**0.80**	0.71	0.76	0.61	**0.79**	0.67	0.78	0.51	**0.80**
MEN-test	0.70	0.79	0.56	**0.81**	0.73	0.74	0.62	**0.79**	0.71	0.79	0.53	**0.80**

Table 3: Spearman's correlation for embeddings before and after the linear transform. All cross-validation (10-fold) experiments have p-value $p < 0.01$.

Results To account for the cross-dimension problem described in the previous section, we design a slightly modified version of the previous experiments. We apply a linear transformation with the weights varying according to which type of word similarity (relatedness or similarity) we are measuring. This allows us to: (1) give more weight to more important relations and (2) compare the representations across different Qualia structure relations.

Using a linear transformation allows us to recover the initial DF representation from its transformed counterpart, which is important in order to maintain the semantic interpretability of DF (i.e. which words are related to t and how). Thus, given DF_t for a term t, we get $DF_t^* = W \times DF_t + b$, which we use in our experiments. The parameters W, b are learnt separately for similarity and relatedness tasks, since different relations and cross-relation comparisons have varying importance for the two tasks. The training objective for the linear transformation is the minimization of the mean squared error between the cosine similarity of the transformed representations and the normalized ground truth similarity score. For fair comparison, we also apply a linear transformation to the baseline $Basis$ by learning parameters W_{basis}, b_{basis} as described above for DF. In our experiments on similarity and relatedness datasets we use 10-Fold cross-validation and report the average performance, while on MEN we use the provided split into training and test data (it is the only dataset with a train/test split).

Our results show that Definition Frames achieve the best performance, compared to any of the baselines. In Table 3 we compare the performance of the Basis embeddings before and after the linear transformation ($Basis$ and $Basis^*$), with the Definition Frames (DF and DF^*). DF^* benefits much more of the dimension weighting and achieves better results compared to $Basis^*$, particularly with GloVe embeddings. Furthermore, we observe that Relatedness datasets (including MEN) gain the greatest advantage from the linear weighting. This lines up with our previous hypothesis, since the relatedness task requires more cross-relation comparisons (DF_{car} vs DF_{wheel}).

Qualitative Analysis One of the distinguishing features of DFs is that they are semantically interpretable. Beyond determining whether two terms are related, we find that DFs can be used to infer *how* they are related. We perform a qualitative analysis on 100 randomly selected terms from the MEN dataset that have high relatedness score (higher than 35 out of 50). The goal of this study is to assess whether we can use the explicit structure of DFs to predict the type of the relation between two terms.

We conduct a Mechanical Turk study, where we present (1) the pair of related words, (2) their corresponding definitions and (3) a Qualia structure relation, in the form of question. We phrase the annotation task as a binary question such as "*Is an aquarium created by a fish?*". We include all possible Qualia structure relations for each of the 100 pairs of related words. We ask three annotators to annotate each sample (1200 questions, each annotated three times, for a total of 3600 annotations).

To identify the most probable relation between two terms t_1 and t_2 using the encoded DF, we conduct a set of row-to-row comparisons. We measure the cosine similarity of each row of DF_{t_1} with $Basis(t_2)$ and vice-versa DF_{t_2} with $Basis(t_1)$. The relation corresponding to the row with highest cosine similarity is taken to be the most probable relation. We test if the relation predicted by the DFs is correct according to humans. By taking the majority vote of the annotations, we find that 77% of the extracted relations are considered valid by the workers. Furthermore, 54% of the relations were considered accurate by all three annotators and the inter annotator percent agreement is 60% over the 1200 relations (more details in Appendix A.3).

5 Conclusion

We propose Definition Frames, a hybrid semantically interpretable representation that is grounded in both lexical semantics and distributed representations. By disentangling the Qualia structure relations, DFs can capture different types of similarity (relatedness and similarity) and achieve improved performance on word similarity tasks. Finally, we demonstrate the explainability of Definition Frames via a human study showing that they provide valid insights on how terms are related. DFs are independent of the distributed representation used as basis. Future work could explore the use of contextual embeddings basis and the benefits of Definition Frames in downstream tasks.

Acknowledgements

This research was partially supported by DARPA grant no HR001117S0017-World-Mod-FP-036 funded under the World Modelers program.

References

Eneko Agirre, Enrique Alfonseca, Keith Hall, Jana Kravalova, Marius Paşca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 19–27. Association for Computational Linguistics.

Luis Espinosa Anke and Steven Schockaert. 2018. Syntactically aware neural architectures for definition extraction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 378–385.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Jean-Louis Binot and Karen Jensen. 1993. A semantic expert using an online standard dictionary. In *Natural Language Processing: The PLNLP Approach*, pages 135–147. Springer.

Guido Boella and Luigi Di Caro. 2013. Extracting definitions and hypernym relations relying on syntactic dependencies and support vector machines. In *51st Annual Meeting of the Association for Computational Linguistics, ACL 2013*, volume 2, pages 532–537. Association for Computational Linguistics (ACL).

Branimir Boguraev and James Pustejovsky. 1990. Lexical ambiguity and the role of knowledge representation in lexicon design. In *Proceedings of the 13th conference on Computational linguistics-Volume 2*, pages 36–41. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Tom Bosc and Pascal Vincent. 2018. Auto-encoding dictionary definitions into consistent word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1522–1532.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 136–145. Association for Computational Linguistics.

Nicoletta Calzolari. 1984. Detecting patterns in a lexical data base. In *10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics*.

Martin S Chodorow, Roy J Byrd, and George E Heidorn. 1985. Extracting semantic hierarchies from a large on-line dictionary. In *Proceedings of the 23rd annual meeting on Association for Computational Linguistics*, pages 299–304. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Manaal Faruqui and Chris Dyer. 2014. Community evaluation and exchange of word vectors at wordvectors.org. In *Proceedings of ACL: System Demonstrations*.

Manaal Faruqui, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1606–1615.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on information systems*, 20(1):116–131.

Kata Gábor, Davide Buscaldi, Anne-Kathrin Schumann, Behrang QasemiZadeh, Haifa Zargayouna, and Thierry Charnois. 2018. Semeval-2018 task 7: Semantic relation extraction and classification in scientific papers. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 679–688.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2009. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 94–99. Association for Computational Linguistics.

Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. 2016. Learning to understand phrases by embedding the dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30.

Douglas B. Lenat. 1995. Cyc: A large-scale investment in knowledge infrastructure. *Commun. ACM*, 38(11):33–38, November.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

George A Miller and Walter G Charles. 1991. Contextual correlates of semantic similarity. *Language and cognitive processes*, 6(1):1–28.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Fons Moerdijk et al. 2008. Frames and semagrams. meaning description in the general dutch dictionary. In *Proceedings of the Thirteenth Euralex International Congress, EURALEX 2008*.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics*, 5:309–324.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition modeling: Learning to define word embeddings in natural language. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237.

James Pustejovsky. 1991. The generative lexicon. *Computational linguistics*, 17(4):409–441.

Herbert Rubenstein and John B Goodenough. 1965. Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

Robert Speer and Catherine Havasi. 2012. Representing general relational knowledge in conceptnet 5. In *LREC*, pages 3679–3686.

Julien Tissier, Christophe Gravier, and Amaury Habrard. 2017. Dict2vec: Learning word embeddings using lexical dictionaries. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 254–263.

Michael Zock and Slaven Bilac. 2004. Word lookup on the basis of associations: From an idea to a roadmap. In *Proceedings of the Workshop on Enhancing and Using Electronic Dictionaries*, ElectricDict '04, pages 29–35, Stroudsburg, PA, USA. Association for Computational Linguistics.

A Appendix

A.1 Relation Retriever performance

In Table 4 we show the performance of the pre-trained Relation Retriever model on ConceptNet data, for all tested models. The performance is evaluated on a held-out test set. We observe that the performance is very high, which is our main motivation to fine-tune on the Qualia annotations of WordNet definitions.

Model	Pr	Re	F1
BiLSTM	97.6	97.7	97.6
BERT BiLSTM	95.1	95.0	95.1
Stacked-BiLSTM	97.6	97.6	97.6
BiLSTM-CNN	97.4	97.6	97.4

Table 4: Relation Retriever on ConceptNet data (held-out test set).

A.2 Ablation Study

We compare the performance of $Basis$ embeddings with Definition Frames where one relation is pruned (*All*-r, when relation r is pruned). In Figure 2 we show the ablation study when we merge the datasets into similarity and relatedness, while in Figure 3, we show the results of the study for each dataset separately.

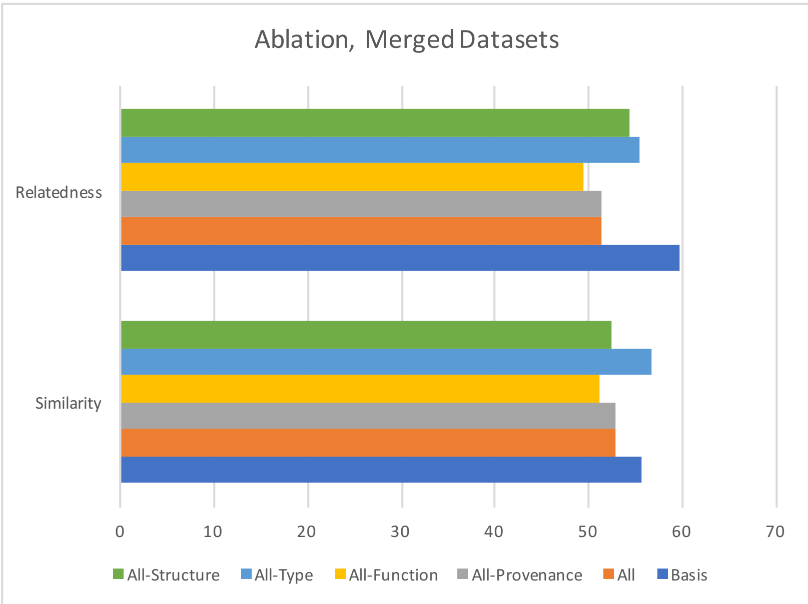

Figure 2: Ablation study for merged datasets.

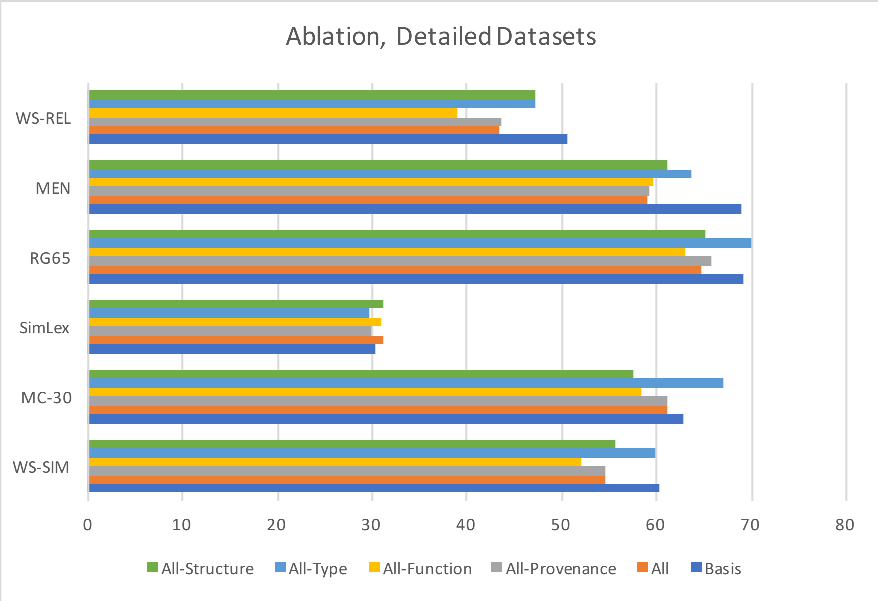

Figure 3: Ablation study for each dataset individually.

A.3 MTurk Study Accuracy

In Table 5, we show the accuracy per relation of the Definition Frames extracted relations, when all three MTurk participants agree.

Qualia	Relation	Agreement %
Formal	IsA	0.43
Constitutive / Structure	PartOf, HasA, MadeOf	0.79
Telic / Function	UsedFor	0.50
Origin / Provenance	CreatedBy	0.25

Table 5: Accuracy per relation.