Enriching Wikidata with Frame Semantics

Hatem Mousselly-Sergieh[1]
Iryna Gurevych[1,2]
[1]UKP Lab, Technische Universität Darmstadt
[2]UKP Lab, German Institute for Educational Research
`https://www.ukp.tu-darmstadt.de`

Abstract

Wikidata is a large-scale, multilingual and freely available knowledge base. It contains more than 14 million facts, however, it is still missing linguistic information. In this paper, we aim to bridge this gap by aligning Wikidata with FrameNet lexicon. We propose an approach based on word embedding to identify a mapping between Wikidata relations, called properties, and FrameNet frames and to annotate the arguments of each relation with the semantic roles of the matching frames. Early empirical results show the advantage of our approach compared to other baseline methods.

1 Introduction

Wikidata (hereafter WD) (Vrandečić and Krötzsch, 2014) is a large-scale, multilingual and freely available knowledge base containing more than 14 million facts. WD entities are directly linked to the corresponding Wikipedia articles. To increase the usability of WD for NLP tasks, we aim at enriching WD with linguistic information by aligning it to the famous lexicon FrameNet (Fillmore et al., 2003).

Several works considered aligning knowledge bases, e.g., Wikipedia with expert-resources like FrameNet and WordNet (Fellbaum, 1998) (refer to (Tonelli et al., 2013; Navigli and Ponzetto, 2012)). However, the focus of these works was on word-sense alignment. That means linking words having the same meaning among different resources. In contrast to previous efforts, we aim to perform the alignment on the relation level. Specifically, we aim to find a mapping between WD facts, e.g. *educated at(Person, University)* and similar structures in expert lexical resources. FrameNet (FN) provides such structure in terms of semantic frames. Briefly, a frame is an abstract description of a situation, e.g. the frame *Education_teaching* and the participants in it, e.g. *Student*, *Teacher* and *Course*.

There are several advantages for such kind of alignment: FN is an essential resource for semantic role labeling (SRL) systems which are usually trained on the annotated corpus that is provided by FN. A crucial problem with such systems is that they are biased towards the domain of that corpus. By linking FN and WD, we could (semi) automatically create another frame-annotated corpus using the links between WD entities and the corresponding Wikipedia articles as well as the alignment between FN and WD. Consequently, the annotated Wikipedia corpus which covers a wide range of domains can be used to improve the performance of SRL systems. As for the knowledge base, in addition to the direct result of enriching WD with linguistic information, the alignments can be used to refine the property structure of WD by inducing new general/specific properties. For instance, the property *killed by* refers to someone (victim) killed by somebody else (killer). However, the property does not distinguish between different kinds of killing, such as execution. In FN such information is already captured through the frames *Execution* and *Killing*, where the former frame inherits from the latter. By aligning *killed by* to both frames, the property *killed by* can refined by introducing a new sub-property: *executed by*.

Our contributions are: (1) a method for extracting semantic representations for WD properties and their arguments, (2) an approach for frame-property as well as role-argument alignment[1] and (3) an experimental evaluation.

The rest of the paper is organized as follows: in

[1] FN-WD alignments: `https://goo.gl/FdhOkO`

Proceedings of AKBC 2016, pages 29–34,
San Diego, California, June 12-17, 2016. ©2016 Association for Computational Linguistics

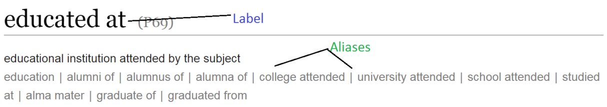
the next Section, a short description of FrmeNet is provided while in Section 3 a method for extracting semantic representation for WD properties and their arguments is presented. Section 4 presents the alignment approach while the results of the experimental evaluation are presented in Section 5. Section 6 discusses related works and a conclusion is provided in Section 7.

2 FrameNet

The main entry in FN is the semantic **Frame** which is a description of a type of event or relation and the participants in it. Each frame consists of a set of semantic roles, called **Frame Elements (FEs)**, which correspond to the participants of the event. Additionally, each frame is associated with a collection of words called **lexical units (LUs)** that evoke that frame. FrameNet provides a corpus of example sentences, in which certain words, named **fillers**, are identified as frame **evoking elements (FEEs)** and annotated with a semantic frame.

3 WD Property Semantic Representation

WD distinguishes between two types of entries: **item** which corresponds to a Wikipedia article and **property** that defines a relation between an item/property and a value, e.g. *educated at(Barack Obama, Columbia University)*. We analyzed the WD data model in order to extract semantic representations for properties and their arguments as a pre-step towards aligning WD with FN. First, we use the notation $p(ARG1, ARG2)$ to refer to a property p and its left-side and right-side arguments, respectively. For each property, i.e., the element p, we extract the following information from the data model: 1) the label of the property and 2) the aliases which are alternative names or loosely speaking synonyms of that property. For example, the following set of semantic representations can generated for the property *educated at* (Figure 1): {educated at, alumni of, college attended, university attended, studied at,...}



educated at (P69)

educational institution attended by the subject

education | alumni of | alumnus of | alumna of | college attended | university attended | school attended | studied at | alma mater | graduate of | graduated from

Figure 1: WD page for the property *educated at*

As for the arguments, we distinguish between two types of semantic representations: **semantic types** and **instances** which we will refer to as ”fillers” in the following discussion. For a given argument, we leverage the structural property *instance of* to obtain the father concept of that argument. Furthermore, we exploits other structural relationships between WD properties, namely *subproperty of*, *inverse of* and *see also* to infer further semantic types about the arguments. Accordingly, the semantic types of the arguments of the related properties are propagated to the arguments of the source property. Take a look at the following instance of the property *father*: *father(George Washington,Augustine Washington)*. Instances of *ARG1* and *ARG2* of this property, i.e., *George Washington* and *Augustine Washington* are linked via *instance of* to the item *human* in WD. Accordingly, we deduce that *ARG1* and *ARG2* are of type *human*. Furthermore, the property *father* is defined as *subproperty* of *relative*. In a similar manner, we extract the semantic types of the property *relative* and use them as descriptors for the arguments of *father*. The same procedure is applied to the properties *see also* and *inverse of* where in the latter case the semantic types are propagated in the reverse order.

For each property in WD, a set of instances can be obtained from the knowledge base. For example, the property *educated at* connects the WD item *Barack Obama* (instance of $ARG1$) to the WD item *Columbia University* (instance of $ARG2$). In analogy to FN, we use the term *fillers* to refer to instances of property arguments. WD provides a large number of such fillers and we use them as further descriptors for property arguments.

4 Towards FrameNet-Wikidata Alignment

Although WD and FN have different objectives, they show considerable overlap in their semantics. Consider the definitions of the frame *Education_teaching* and the property *educated at*:

- **Education_teaching**: *This frame contains words referring to teaching and the participants in teaching. A Student comes to learn either about a Subject; a Skill; a Precept; or a Fact as a result of instruction by a Teacher.*

- **student of**: *person who has taught this person.*

Although the definitions have different granularity, their overlap is obvious. Moreover, the arguments $ARG1$ and $ARG2$ of *student of* (with the semantic types *student* and *teacher*, respectively), represent direct correspondences to the FEs *Student* and *Teacher*, respectively. However, the conceptual differences implies that the alignment between frames and properties is rather many-to-many than one-to-one. Additionally, properties are more specific than frames in the sense that they describe a single fact rather than a situation. Hence, a partial alignment between property arguments and FEs is natural.

4.1 Property-Frame Alignment

First, we aim to align WD properties with FN frames. For this purpose, we create for each property a context based on its label and aliases (refer to Section 3). Similarly, we create a context for each frame based on its lexical units and frame label.

In contrast to the rich frame context (each frame is associated with 13 lexical units on average), property context is rather poor. This is because a considerable part of WD properties has few to no aliases. Therefore, we expand the property context with additional words based the technique of word embedding (Mikolov et al., 2013b). Word embedding is a technique for representing words as vectors of real numbers in a low-dimensional space. It has gained much attention recently and has been successfully applied to a wide range of semantic tasks (Faruqui and Dyer, 2014). (Levy and Goldberg, 2014) presented a word embedding approach in which the context of a given word is created based on the dependency graph of that word over large collection of sentences. According to this approach, words with similar functionality, such as co-hyponyms lay close to each other in the embedding space. This type of embedding is good candidate for our case because we assume that words of similar functionality would evoke the same frame. Therefore, we use the pre-calculated word vectors provided by (Levy and Goldberg, 2014) to expand the context of WD properties. First, we identify for each label and alias (if available) a set of words that are close to them in the dependency-embedding space. Next, we combine the embedding vectors by summing them to obtain a single embedding vector for each property context. We also experimented with different combination methods, e.g. averaging, multiplication and subtraction, however, the sum led to the best results.

Similarly, we create for each frame context an embedding vector by looking up the corresponding words in the same embedding space and summing the identified embedding vectors.

Finally, the property-frame alignments are determined based on the cosine similarity between the final embedding vectors of the two contexts. Figure 2 illustrates the described alignment procedure.

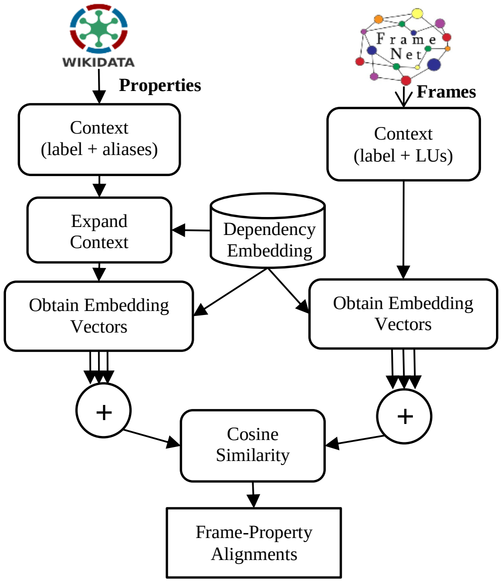


Figure 2: Property-Frame alignment workflow

4.2 Argument-FE Alignment

After identifying property-frame correspondences, property arguments are mapped to FEs as follows (Figure 3):

1) Creating Argument/FE Context:
Regarding property arguments, we apply the procedure described in Section 3 to create two contexts for each argument: *semantic type* and *filler* contexts. Similarly, we create for each FE two contexts: 1) semantic type context which consists of the label and the semantic type of that FE as defined in FN and 2) filler context which contains the headwords of the fillers of that FE which were obtained from the FN annotated corpus according to (Bauer et al., 2012).

2) Generating Word Embedding Vectors:
Next, the embedding vector for each word in the argument/FE context are retrieved from a word embedding space that was trained on the Google News

dataset as provided by the word2vec framework (Mikolov et al., 2013a). We chose this embedding space due to its high coverage of three million words and phrases. Indeed, phrases are crucial in our case, especially, since the majority of argument fillers correspond to named entities. Subsequently, the embedding vectors are summed to produced one final vector per context, i.e., one for the semantic type context and one for the filler context.

3) Calculating Argument-FE Similarities:
In this step, the pairwise similarity between each argument a and FE e of a matched frame-property pair is calculated. The similarity is based on a combination of two scores, i.e., the cosine similarities between the semantic type contexts and the filler contexts of a and e, respectively:

$$Sim(a, e) = \alpha S(V_a, V_e) + (1 - \alpha) S(W_a, W_e) \quad (1)$$

V_a/W_a, V_e/W_e are the combined embedding vectors of the semantic type/filler contexts of a and e, respectively, S is the cosine similarity and $\alpha \in [0, 1]$ is a weighting parameter that is used to tune the effect of the semantic type/filler contexts on the final similarity. Setting α to 0.5 leads to a equal effect of both contexts, $\alpha = 1$ ignores the filler contexts while $\alpha = 0$ eliminates the semantic type contexts from the similarity calculation.

The similarity scores are then used to determine the final alignments. Here, we ensure that the final alignments satisfy two constraints: 1) each argument is aligned to at most one FE and 2) each FE is aligned to at most one argument.

5 Evaluation

We created a gold standard from a sample of 130 WD properties. For each property, two annotators were provided with a list of 7 candidate frames on average and had to answer the question whether a property-frame pair is a match or not based on the corresponding definitions and an example per property/frame. The inter-annotator agreement according to Cohen's κ was 0.65. After removing the disparagement pairs, the gold standard contained 785 property-frame pairs with 279 positive and 506 negative alignments, respectively. For the proportion of positive alignments the same annotators also aligned the arguments and FEs. The final set contains 411 argument-FE alignments.

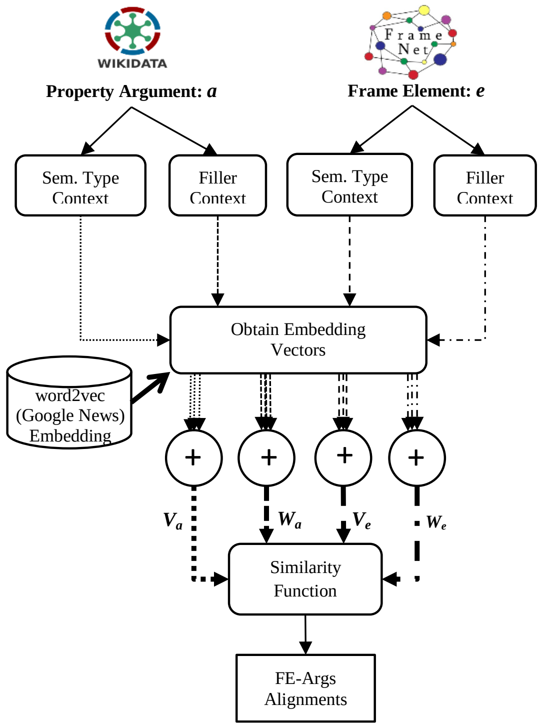


Figure 3: Argument-FE alignment workflow

5.1 Results: Property-Frame Alignment

The alignment approach was applied on FN version 1.5 which contains 1,019 frames and WD dump of 28/9/2015 which contains 1,745 properties. After filtering properties that describes identifiers (e.g. the property *GND identifier*) or structural relationships (e.g. *subproperty of*, *facet of*), we were able to align 638 properties (37% of the total WD properties) to a total of 380 unique frames (37% of the total frames).

We compared the performance of our method to other baselines. In the first baseline (BL1) the alignment is determined based on the lexical overlap between the frame and property contexts without expansion. The second baseline (BL2) expands the property context with words from the most frequent WordNet synsets instead of using the embedding space. Next, the embedding vectors of the expanded property context and frame context are summed and the cosine similarity is applied on the final context vectors.

For each property the top two matching frames were identified and precision, recall and f1-measure were reported (Table 1). The results show that enriching the context of the property with further words either from WordNet or using a given embedding space leads to better results compared to BL1. Moreover, expanding the property context using dependency-based embeddings (our method) outperforms WordNet based expansion (BL2).

Method	P	R	F1
BL 1	0.45	0.44	0.45
BL 2	0.65	0.68	0.66
Our Method	**0.70**	**0.72**	**0.73**

Table 1: Performance of frame-property alignment

5.2 Results: Argument-FE Alignment:

This task was evaluated by measuring the accuracy of the matching as the proportion of correctly aligned property arguments and taking the average. We experimented with different values of α (Equation 1). The experiments showed that the filler context has higher influence on the accuracy than the semantic type context (best results are obtained with $\alpha = 0.35$). Accordingly, we can conclude that the semantic types are less discriminative than the fillers. For example, the two arguments of the property *killed by* share the semantic types *human* and *person*, thus, it is impossible to determine which argument represents the victim and which one represents the killer. However, by using the fillers a better distinction can be made.

We also compared our approach to other baselines which use the filler as well as the semantic type contexts as input, however, without applying the described word embedding approach. We investigated three similarity measures: the lexical overlap, Jaccard similarity and the cosine similarity between the context vectors. Our approach outperforms the baselines (Table 2) and the results confirms the advantage of using word embedding for this task.

	ARG 1	ARG 2	AVG
	Accuracy	Accuracy	Accuracy
Overlap	0.55	0.56	0.56
Cosine	0.51	0.62	0.57
Jaccard	0.53	0.63	0.58
Our Method	**0.70**	**0.68**	**0.69**

Table 2: The accuracy of argument-FE alignments

6 Related Work

The problem of aligning expert lexical resources in order to increase their coverage was the topic of several research efforts (Shi and Mihalcea, 2005; Chow and Webster, 2007; Johansson and Nugues, 2007; De Cao et al., 2008; Lacalle et al., 2014). Another line of research considered aligning community-created resources like Wikipedia and Wiktionary to lexical resources like FrameNet. (Tonelli and Giuliano, 2009; Tonelli et al., 2013) presented an approach for extending FN by linking its LUs to Wikipedia articles using supervised WSD. (Hartmann and Gurevych, 2013) presented an approach for linking FN with Wiktionary in order to build a FrameNet-like resource for German. While our work consider the alignment on the relation level, the mentioned efforts focus on extending the coverage of FN by inducing new LUs using word-sense alignment techniques. In fact, the problem of aligning FN frames with knowledge base relations is new. An initial attempt with a similar goal as ours was presented by sar-graph (Krause et al., 2015). *sar-graph* is a graph that connects different contractions of a given relation. The nodes correspond to words or arguments of that relation and are labeled with lexical, syntactic and semantic information. The authors presented initial ideas for linking sar-graphs with FN using valency and phrase patterns and claimed that such a connection would allow linking FN frames with sar-graph relations. Compared to our work, where a concrete solution is presented, the mapping between sar-graph relations and FN is still in its early stage.

7 Conclusion and Future Work

We presented an approach for aligning WD with FN which addresses two tasks: frame-property mapping as well as mapping property arguments to FEs of the matching frames. We presented a simple but effective alignment approach based on the technique of word embedding. In future work, we will evaluate the advantage of the created alignments in the context of semantic role labeling.

References

Daniel Bauer, Hagen Frstenau, and Owen Rambow. 2012. The dependency-parsed framenet corpus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uur Doan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Ian C Chow and Jonathan J Webster. 2007. Integration of linguistic resources for verb classification: Framenet frame, wordnet verb and suggested upper merged ontology. In *Computational Linguistics and Intelligent Text Processing*, pages 1–11. Springer.

Diego De Cao, Danilo Croce, Marco Pennacchiotti, and Roberto Basili. 2008. Combining word sense and usage for modeling frame semantics. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, pages 85–101. Association for Computational Linguistics.

Manaal Faruqui and Chris Dyer. 2014. Community evaluation and exchange of word vectors at wordvectors.org. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Baltimore, USA, June. Association for Computational Linguistics.

Christiane Fellbaum, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Charles J Fillmore, Christopher R Johnson, and Miriam RL Petruck. 2003. Background to framenet. *International journal of lexicography*, 16(3):235–250.

Silvana Hartmann and Iryna Gurevych. 2013. Framenet on the way to babel: Creating a bilingual framenet using wiktionary as interlingual connection. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*, volume 1, pages 1363–1373, Stroudsburg, PA, USA, August. Association for Computational Linguistics.

Richard Johansson and Pierre Nugues. 2007. Using wordnet to extend framenet coverage. In *Building Frame Semantics Resources for Scandinavian and Baltic Languages*, pages 27–30. Department of Computer Science, Lund University.

Sebastian Krause, Leonhard Hennig, Aleksandra Gabryszak, Feiyu Xu, and Hans Uszkoreit. 2015. Sar-graphs: A linked linguistic knowledge resource connecting facts with language. In *Fourth Workshop on Linked Data in Linguistics: Resources and Applications (LDL-2015) at ACL-IJCNLP 2015*. ACL.

Maddalen Lopez De Lacalle, Egoitz Laparra, and German Rigau. 2014. Predicate matrix: extending semlink through wordnet mappings. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Omer Levy and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 302–308.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Roberto Navigli and Simone Paolo Ponzetto. 2012. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, December.

Lei Shi and Rada Mihalcea. 2005. Putting pieces together: Combining framenet, verbnet and wordnet for robust semantic parsing. In *Computational linguistics and intelligent text processing*, pages 100–111. Springer.

Sara Tonelli and Claudio Giuliano. 2009. Wikipedia as frame information repository. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 276–285, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sara Tonelli, Claudio Giuliano, and Kateryna Tymoshenko. 2013. Wikipedia-based wsd for multilingual frame annotation. *Artificial Intelligence*, 194:203–221.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, September.