Universal Schemas Using Shortest Dependency Paths for Free Word Order Languages

Jiho Kim, Sangha Nam, and Key-Sun Choi

Korea Advanced Institute of Science and Technology, Republic of Korea
{hogajiho, sangha.nam, kschoi}@kaist.ac.kr

Abstract. Universal schemas are a remarkable approach for solving relation extraction, in which new facts are extracted by jointly learning latent feature vectors of entity pairs and relation types through matrix factorization models. However, in free word order languages where surface form predicates do not fully reveal the characteristics of a sentence, universal schemas cannot be constructed in the same manner. Therefore, in this study, we introduce a novel expansion of universal schemas, dependency-path-based universal schemas. Our model uses shortest dependency paths and entity types instead of surface form predicates. For verification of our model, we constructed and evaluated a universal schema in Korean with a combination of Korean DBpedia and Korean Wikipedia distant supervision data.

Keywords: Universal Schemas · Relation Extraction · Shortest Dependency Paths

1 Introduction

Relation extraction (RE) is a core step in various natural language processing applications, and can be defined as the task of finding ontological relations between two target entities in a sentence. The concept of universal schemas [5] was proposed as a considerable approach in RE. A universal schema can be constructed using a combination of all usable knowledge bases (KBs) and natural language text, then expressed through a huge matrix with entity tuples as rows and relations as columns. The novelty of universal schemas comes from avoiding pre-labeled datasets by using surface form predicates as a source for relation types, and mutually supporting both unstructured and structured data.

In English, surface form predicates between two entities can usually serve as relations [1]. However, in free word order languages, the extraction of surface pattern relations between entities is difficult owing to flexible word orders. Without the presence of language-specific surface form predicate extractors, universal schemas cannot be built in these languages. Even though several algorithmic extensions have been studied [4, 6], none have solved the dependency on surface form predicates.

In this study, we propose a novel expansion of universal schemas: dependency path based universal schemas (DPUSs). Shortest dependency paths (SDPs) illustrate the syntactic structure between entities according to a sequence of directed

binary grammatical relations. Unlike surface form predicates, grammatical dependency is a common feature in all languages. We also injected entity types into the matrix, expecting the effect of narrowing of the range of relations that the model should predict. To verify that our model shows a decent performance, we constructed two DPUSs (baseline/extended) on a Korean dataset and measured the average precision (AP) for each ontological relation.

2 Methods

We first built a matrix with entity tuples as rows and a combination of KB relations, SDPs, and entity types as columns. For all observed and unobserved facts, we filled in 1 and 0, respectively in the corresponding slot. We extracted SDPs between the target entities, including directions of each dependency. Figure 1 illustrates the filling of the DPUS matrix based on raw text. Our model mainly employs the *nfe* matrix factorization model from [5], which showed the best performance among the proposed models. The embedding vector dimensions were optimized using the leave-one-out cross-validation. To ensure consistency of slot values, we normalized each row on a scale of 0 to 1.

3 Experiments

3.1 Dataset and Evaluation

For the baseline, we built a DPUS by using Korean DBpedia [2] as the KB and Korean Wikipedia distant supervision data as raw text. The distant supervision data is obtained from Korean Wikipedia documents, which were filtered by Korean DBpedia triples. Specifically, we extracted sentences, each of which contained two entities with a relation in Korean DBpedia. We rejected sentences with SDP of length longer than three because these SDPs are rarely observed in our dataset. For our target relation set, we selected the most frequent 35 predefined KB relations, each with more than 200 entity tuples in the distant supervision data. Then, we chose entity tuples with two or more SDPs in the distant supervision data. We finally obtained 12,360 entity tuples, 35 KB relations, and 348 frequent SDPs. The size of the baseline DPUS matrix was 12360×383.

For the extended model, we added coarse-grained entity types extracted by Named Entity Recognition (NER) of each entity pair to the baseline DPUS matrix. We used the following entity type categorization: artifacts (e.g., TV programs, books, and movies), data/time, locations (e.g., countries, cities, and towns), organizations (e.g., governments, public corporations, and companies), persons, and others. For each entity tuple, we added a new slot indicating the entity types in the tuple. A total of 15 new columns were added to the baseline matrix, extending the size of the extended matrix to 12360×398.

For evaluation, we performed a ten-fold experiment to measure APs for each KB relation over all the entity tuples in the test set. Then we measured MAP and WMAP for the performance of the whole model. MAP is simply the average of

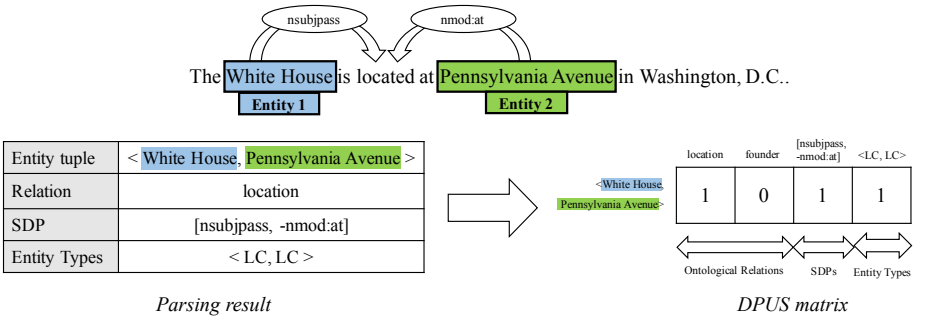

Fig. 1. Example of building DPUS matrix. The two entities with an ontological relation location are connected with the SDP [nsubjpass, -nmod:at], and both the entities are of type location (LC). We convert these parsing results into the DPUS matrix by filling the corresponding slots. Note that the "-" symbol in SDP implies inverse dependency.

APs, and WMAP is the weighted version of MAP in which each AP is weighted by the number of entity tuples in each relation. MAP and WMAP are shown by a previous work [3] to be robust and stable metrics for evaluating classification models.

3.2 Results and Discussion

The experimental results of all KB relations are listed in Table 1. Winners of each relation are marked in bold. The baseline DPUS showed performances of 0.62 (MAP) and 0.64 (WMAP), while the extended version of DPUS showed performances of 0.72 (MAP) and 0.75 (WMAP), respectively. Thus, the performance shows great improvement after the injection of entity types. This is because for each entity tuple, the model can consider a significantly fewer set of relations when the entity types are decided.

Even though the extended DPUS shows better overall performance, the baseline method performs better in some relations, such as *birthPlace* and *writer*. As most of these relations are represented by unique SDPs, the injection of entity types seems to confuse the model.

4 Conclusion

In this study we introduced DPUSs, which are a novel expansion of universal schemas. DPUSs are built by using SDPs and entity tuple types instead of surface form predicates. SDPs contain information about the syntactic structure of a sentence, which makes SDPs a suitable alternative of surface form predicates. Entity tuple types greatly improve performance by reducing the relation candidates for the model. Our model has higher potential in terms of future development and versatility because DPUS is based on globally consistent features rather than language specific features.

Relation	#	Baseline	+NER	Relation	#	Baseline	+NER
country	166	0.49	**0.91**	channel	25	0.57	**1.00**
team	73	0.81	**0.91**	currentMember	25	0.91	**0.94**
isPartOf	67	0.57	**0.83**	parent	24	**0.65**	0.51
birthPlace	63	**0.80**	0.73	city	22	0.37	**1.00**
deathPlace	51	**0.66**	0.51	owner	22	0.69	**0.76**
location	50	0.46	**0.76**	operator	20	0.75	**0.76**
associatedBand	47	**1.00**	0.93	activeYearsEndYear	20	0.79	**0.76**
associatedMusicalArtist	47	**1.00**	0.95	bandMember	19	**0.78**	0.70
club	46	**0.83**	0.68	spouse	18	0.33	**0.64**
nationality	39	**0.62**	0.33	artist	18	0.58	**0.79**
predecessor	33	**0.64**	0.53	capital	17	0.53	**0.84**
writer	33	**0.71**	0.56	managerClub	14	**0.32**	0.23
director	30	0.81	**0.87**	routeStart	14	0.60	**0.66**
producer	29	**0.60**	0.52	routeEnd	14	0.53	**0.66**
region	27	**0.26**	0.09	activeYearsStartYear	14	0.67	**0.88**
position	27	0.38	**0.94**	deathYear	12	0.58	**0.74**
successor	27	0.50	**0.66**	author	12	0.38	**0.81**
league	27	0.59	**0.95**				
MAP		0.62	**0.72**	**WMAP**		0.64	**0.75**

Table 1. AP and WMAP of of two DPUS models for 35 KB relations. The "#" column indicates the number of triples in the test set with the corresponding relation. The "*+NER*" column indicates the results of the extended DPUS with entity types injected.

Acknowledgement

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (2017-0-01780, The technology development for event recognition/relational reasoning and learning knowledge based system for video understanding)

References

1. Etzioni, O., Banko, M., Soderland, S., Weld, D.S.: Open information extraction from the web. Communications of the ACM **51**(12), 68–74 (2008)
2. Kim, E.k., Weidl, M., Choi, K.S., Auer, S.: Towards a korean dbpedia and an approach for complementing the korean wikipedia based on dbpedia. OKCon **575**, 12–21 (2010)
3. Manning, C.D., Raghavan, P., Schütze, H., et al.: Introduction to information retrieval, vol. 1. Cambridge university press Cambridge (2008)
4. Neelakantan, A., Roth, B., Mc-Callum, A.: Compositional vector space models for knowledge base inference. In: 2015 AAAI spring symposium series (2015)
5. Riedel, S., Yao, L., McCallum, A., Marlin, B.M.: Relation extraction with matrix factorization and universal schemas. In: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 74–84 (2013)
6. Yao, L., Riedel, S., McCallum, A.: Universal schema for entity type prediction. In: Proceedings of the 2013 workshop on Automated knowledge base construction. pp. 79–84. ACM (2013)