Reconciling Event-Based Knowledge through RDF2VEC

Mehwish Alam[1], Diego Reforgiato Recupero[2,3], Misael Mongiovi[2], Aldo Gangemi[1,2], Petar Ristoski[4]

1. LIPN, Université Paris 13, France, 2. ISTC-CNR, Rome, Catania, Italy, 3. University of Cagliari, Italy, 4. University of Mannheim, Germany.

Abstract. The reconciled knowledge graphs are typically used for multi-document summarization, or to detect knowledge evolution across document series. This paper focuses on reconciling knowledge graphs generated from two text documents about similar events described differently. Our approach employs and extends MERGILO, a tool for reconciling knowledge graphs extracted from text, using word similarity and graph alignment. Complete semantic representation of events are generated using FRED, a semantic web machine reader, jointly with Framester, a linguistic linked data hub represented using a novel formal semantics for frames. Event-reconciliation is mainly performed via similarities based on the graph structure of frames using RDF2Vec graph embeddings, and the subsumption hierarchy of semantic roles as defined in Framester. Our approach is evaluated over a coreference resolution task.

Keywords: Knowledge Reconciliation, Event Reconciliation, Frame Embeddings, Frame Similarity, Role Similarity, Role Embeddings, Framester.

1 Introduction

This study targets the problem of knowledge reconciliation (KR) [18] from the perspective of events. KR is useful in providing a combination of multiple graphs generated by multiple texts describing the same event. This merged graph provides a graph based summary of multiple texts which is more easily comprehensible by users and machines and usable by the algorithms providing interactive exploration of graphs/text analytics through visualization methods.

MERGILO [18] is a tool for reconciling knowledge graphs extracted from text, it first computes the word similarity between the node labels and then performs graph alignment over the complete graphs. When different verbs denote similar events and different agents play slightly different roles, the string matching techniques as introduced in MERGILO might not be appropriate in the KR process. For overcoming this limitation we use *Frame Semantics* which describes a situation in the text with the help of frames and roles. For identifying frames and semantic roles of entities in a text we use FRED [12], a machine reader which generates event-centered knowledge graphs from two different texts.

Then, the similarity between these events is computed by calculating the similarity between the corresponding FrameNet [2] frames and semantic roles (frame elements). We adapt WordNet [10] similarity measures [4] to frames and roles and vector based similarities using the FrameNet graph and the subsumption hierarchy of roles as defined in Framester [11]. We follow the approach `RDF2Vec` [23] to generate graph based *frame embeddings*, used to calculate the semantic similarity between frames. It uses graph mining algorithms such as graph walks and graph kernels to traverse the graph for generating sequences, which are then fed to neural model for generating its vector representations. An evaluation on Cross-document coreference resolution shows significant improvement over the baseline.

The rest of this paper is structured as follows. Section 2 lists the data sources, resources and tools we have adopted in our methodology. Section 3 includes state of the art work. Then, Section 4 gives some details of MERGILO and its functionalities and explains how frame semantics have been employed for improving MERGILO. Section 5 shows a precision-recall analysis for the presented approach on the EECB dataset. Finally, Section 6 concludes the paper with discussions, remarks and highlights some future directions.

2 Role Oriented Resources

FrameNet [2] contains *frames*, which describe a situation, state or action. Each frame has *frame elements* usually consisting of agent, patient, time and location and are also known as *semantic roles*. Each frame can be evoked by *Lexical Units (LUs)* belonging to different parts of speech. These LUs can be nouns, verbs, adjectives and adverbs representing closely related sets of meanings. For example, in the frame *Conquering* the argument for the role *Conqueror* overtakes the argument of the role *Theme* where the theme loses its autonomy. Such constructs describing the situation of conquering or invasion are referred to as *frame elements* and the LUs such as *conquer*, *overtake* etc. are example words, typically used to denote conquering situations in text. In the example bellow, *The Spaniards* is the argument of the role *Conqueror* and *Incas* is the argument of the role *Theme* and *conquered* is the LU evoking the frame.

$$[\textit{The Spaniards}]_{Conqueror}\ [\textit{conquered}]_{Lexical\ Unit}\ [\textit{the Incas}]_{Theme}. \tag{1}$$

Framester [11] is a large RDF[1] knowledge graph (currently including about 30 million RDF triples) acting as a hub between FrameNet, WordNet, VerbNet [14], BabelNet [19], Predicate Matrix [6], etc. Framester uses a mapping between WordNet, BabelNet, VerbNet and FrameNet at its core using detour based approach, expands it to other linguistic resources transitively. It further links these

[1] `https://www.w3.org/TR/rdf11-primer/`

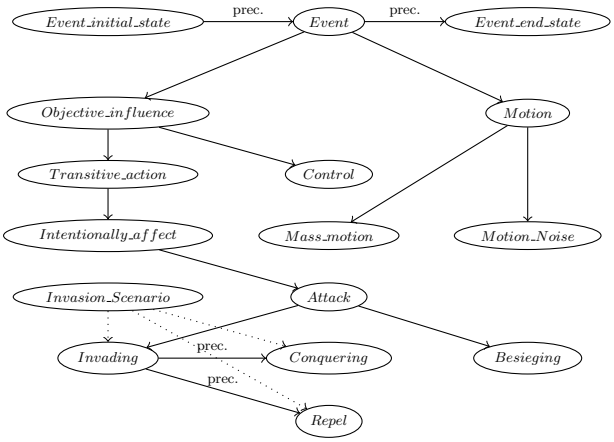

Fig. 1: A part of FrameNet graph. "prec." represents the relation "precedes", dotted lines represent "SubFrame" relation and solid lines represent the "Inheritance" relation as defined in FrameNet.

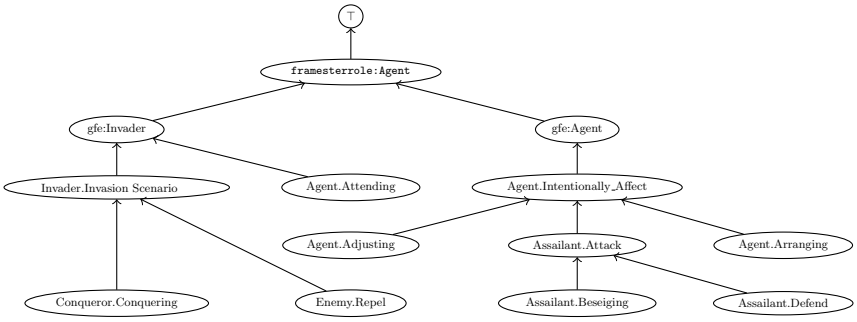

Fig. 2: A part of Subsumption Hierarchy with FrameNet and Framester Roles.

resources to important ontological and linked data resources such as DBpedia, YAGO, DOLCE-Zero [20], schema.org etc.

Framester keeps the original FrameNet graph where the nodes represent the FrameNet frames and the edges represent different semantic relations between the frames i.e., *Inheritance, SubFrame, CausativeOf* etc. Figure 1 shows a part of FrameNet graph. Framester also contains a new subsumption hierarchy of semantic roles (i.e., frame elements) and added generic roles on top of the frame specific roles. Figure 2 shows a part of the Framester role hierarchy associated with the framester role *agent*.[2]

[2] The prefixes for `http://www.ontologydesignpatterns.org/ont/framenet/abox/gfe/` and `http://www.ontologydesignpatterns.org/ont/framester/data/framesterrole.ttl#` are `gfe:` and `framesterrole:` respectively.

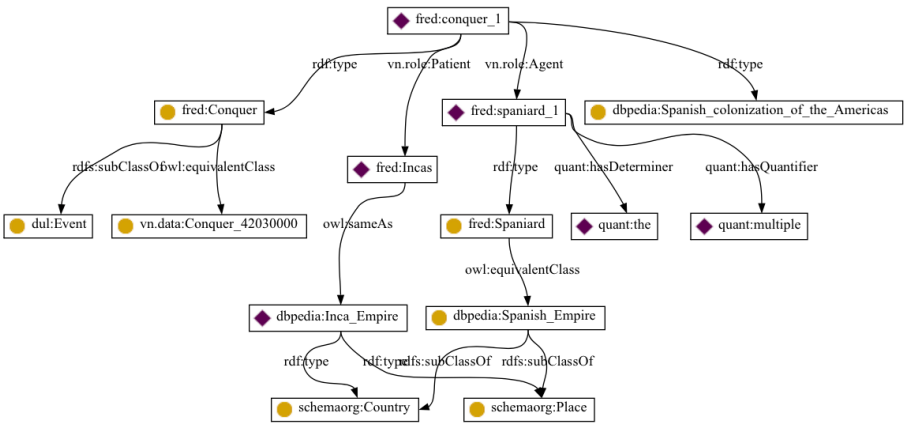

Fig. 3: FRED Knowledge Graph for *Example 1*

FRED [12] [3] is a machine reader which generates ontological structure from natural language text using Discourse Representation Theory (DRT), frame semantics and Ontology Design Patterns. FRED uses Boxer,[4] an open source tool for deep parsing of natural language using Combinatory Categorial Grammar (CCG) and produces event-based, semantic representations of natural language. The Discourse Representation Structures (DRS) produced by Boxer use VerbNet thematic roles. These functionalities implemented in FRED help in the event detection task for our method.

3 State of the Art

Approaches for integrating knowledge include cross-document coreference resolution (when knowledge is represented as text documents) and ontology matching (when knowledge is in a machine-readable form). Cross-document coreference resolution aims at associating mentions about a same entity (object, person, concept, etc.) across different texts [8]. When extracted entities are events, the problem changes to resolution of event coreference across documents [3]. The authors in [16] jointly model named entities and events. Clusters of entities and event mentions are constructed and merged accordingly to a similarity threshold based on linear regression. Then, information flows between entity and event clusters through features that model semantic role dependencies. The system handles nominal and verbal events as well as entities, and the joint formulation allows information from event coreference to help entity coreference, and vice-versa. A rich overview of ontology matching methods is provided by [9]. Relevant work includes [24] that leverages the interplay between schema and

[3] `http://wit.istc.cnr.it/stlab-tools/fred`
[4] `https://github.com/valeriobasile/candcapi`

instance matching. Similarly, [15] shows a greedy iterative algorithm for aligning knowledge bases with millions of entities and facts. These approaches are characterised by the preferred large size of the ontologies/datasets treated (for best performance), which is rarely (probably never) derived from text sources. MERGILO, as other knowledge integration tools [15], employs graph alignment, a more general and widely studied problem [26]. Note that all these approaches are connected and related to the classical graph matching problem [22]. We address this problem from the perspective of events, by taking advantage of frame embeddings i.e., the vector representations of linguistic frames and semantic roles.

Recently, word embeddings have been used in variety of Information Retrieval and Natural Language Processing applications. One recent application is used for generating vector representations of word senses [13] and then these vector representations are used for improving the results of word similarity and word analogy tasks based on BabelNet word senses formally known as `SensEmbed`. [5] apply Frame Semantics and Distributional Semantics for slot filling in Spoken Dialogue System. In [27], the authors use Word and Frame Embeddings for generating categories of annoying behaviors where each category contains a set of words specific to that category. The frame embeddings are generated using 3.8 million tweets tagged by FrameNet frames using SEMAFOR. By contrast, in this study we are using graph-based Frame and Role Embeddings.

4 Event-Based Knowledge Reconciliation

Consider the two sentences: **Sent1:** *"The Spaniards conquered the Incas."* and **Sent2:** *"The Incas were attacked by the Spaniards."* They are describing the same event in the past using different words i.e., event of an attack or an invasion from *Spaniards* to *Incas*. Figure 3 shows the FRED graph of *Sent1*. Given two such knowledge graphs, MERGILO first performs graph compression by merging the nodes in the same graph. The two compressed graphs are aligned by establishing a 1-1 correspondence between the nodes of the two graphs by maximizing a score function, which combines the similarity between aligned nodes and the similarity between aligned edges. In such a case, the similarity between "conquered" and "attacked" is not effective since the word similarity is low, although in this context such words describe the same event.

For computing similarity between two nodes containing verb senses, the verb senses are first mapped to frames using Framester mappings. For example, in Figure 3 s_1 = *vn.data* : *Conquer_42030000* and for *Sent2* we have s_2 = *vn.data* : *Attack_33000000*. According to Framester mappings, we obtain $s_1 \rightarrow \{Conquering\}$ and $s_2 \rightarrow \{Attack\}$. These nodes are replaced by their corresponding frames. The edges containing the VN-roles are mapped to FN-roles. For example, in Figure 3, the verb sense `vn.data`[5]`:Conquer_42030000` evokes the roles `vn.role:Agent` and `vn.role:Patient` which are mapped to `fe:Conqueror.conquering` and `fe:Theme.conquering` respectively.

[5] `prefix vn.data: http://www.ontologydesignpatterns.org/ont/vn/vn31/data/`

In the sentence in Figure 2, the roles evoked by the verb sense `vndata:Attack_33000000` are `vndata:Agent` and `vndata:Theme`. The Framester mappings contains the following records for these roles:

```
vndata:Agent.conquer_42030000 skos:closeMatch fe:Conqueror.conquering .
vndata:Patient.conquer_42030000 skos:closeMatch fe:Theme.conquering .
vndata:Agent.attack_33000000 skos:closeMatch fe:Assailant.attack .
vndata:Theme.attack_33000000 skos:closeMatch fe:Victim.attack
```

Then the similarities are computed in two ways: *(i)* by considering the taxonomical structure imposed by the "inheritance" relation represented as `fnschema`[6]`:inheritsFrom` in Framester using Path Similarity, Wu-Palmers Similarity, Leacock-Chodorow Similarity; *(ii)* using Frame Embeddings.

Frame Embeddings using RDF2Vec: To learn latent numerical representation of the frames and roles in the FrameNet graph, we follow the RDF2Vec approach. First we transform the graph into a set of sequences of entities, which is then fed into a neural language models, resulting into vector representation of all the nodes in the graph in a latent feature space.

To convert the graph into a set of sequences of entities we use two approaches, i.e., graph walks and Weisfeiler-Lehman Subtree RDF Graph Kernels. *(i) Graph Walks:* given a graph $G = (V, E)$, for each vertex $v \in V$, we generate all graph walks P_v of depth d rooted in vertex v. To generate the walks, we use the breadth-first algorithm. In the first iteration, the algorithm generates paths by exploring the direct outgoing edges of the root node v_r. In the second iteration, for each of the previously explored edges, the algorithm visits the connected vertices. The final set of sequences for the given graph G is the union of the sequences of all the vertices $P_G = \bigcup_{v \in V} P_v$. *(ii) Graph Kernels:* it computes the number of sub-trees shared between two or more graphs by using the Weisfeiler-Lehman [7] test of graph isomorphism. This algorithm creates labels representing subtrees.

Once the set of sequences of entities is extracted, we build a word2vec model. Word2vec is a particularly computationally-efficient two-layer neural net model for learning word embeddings from raw text. There are two different algorithms, the Continuous Bag-of-Words model (CBOW) and the Skip-Gram model. The CBOW model predicts target words from context words within a given window, while the skip-gram model does the inverse. Once the training is finished, the cosine similarity is computed between two frames and roles.

5 Evaluation

The experiments were conducted for the task of Cross-document Coreference Resolution (CCR) on RDF graphs, which focuses on associating RDF nodes about a same entity (object, person, concept, etc.) across different RDF graphs generated from text. The data set used for the experimentation was obtained

[6] `prefix fnschema: http://www.ontologydesignpatterns.org/ont/framenet/tbox/`

by the EECB data set which specifies coreferent mentions (text fragment). Our dataset was obtained by generating RDF graphs using FRED and associating text mensions to graph nodes by manual annotations. The framework is built on top of the original MERGILO code, which was released as a Python tool[7]. IBM ILOG CPLEX 12.6.1 was used for solving the Integer Linear Program and the experiments were conducted on a MacOS server with 6-Core Intel Xeon E5 3.50GHz and 64GB of RAM. We used the following metrics for evaluation: (i) MUC [25]: Link-based metric that quantifies the number of merges necessary to cover predicted and gold clusters; (ii) B^3 [1]: Mention-based metric that quantifies the overlap between predicted and gold clusters for a given mention; (iii) CEAFM (Constrained Entity Aligned F-measure Mention-based) [17]: Mention-based metric based on a one-to-one alignment between gold and predicted clusters; (iv) CEAFE (Constrained Entity Aligned F-measure Entity-Based) [17]: Entity-based metric based on a one-to-one alignment between gold and predicted clusters; (v) BLANC (Bilateral Assessment of NounPhrase Coreference) [21]: Rand-index-based metric that considers both coreference and non-coreference links.

For the current evaluation, MERGILO was considered as a baseline. Table 1 shows the results for the baseline method, the Wu-Palmer's similarity, the Path similarity, the Leacock-Chodorow similarity and the results for cosine similarity using *(i)* graph walks and *(ii)* graph kernels with FrameNet roles respectively. Here `Frame2Vec` refers to the vector representations generated for FrameNet frames and `Role2Vec` refers to the vector representations generated for frame elements i.e., semantic roles.

For the first approach with graph walks, for each entity in the FrameNet graph 200 and 500 random walks were generated, each of depth 4 and 8. For each entity in the subsumption hierarchy of roles we generate 400 random walks with depth 4. For the Weisfeiler-Lehman algorithm, we use $h = 2$ iterations and subgraph depth $d = 2$, and after each iteration of the algorithm we extract all walks for each entity with the same depth. We use these sequences to build both CBOW and Skip-Gram models with the following parameters: window size = 5; number of iterations = 10; negative sampling for optimization; negative samples = 25; with average input vector for CBOW. We experiment with 200 and 500 dimensions for the entities' vectors.

The results clearly indicate that each model used for graph walks and graph kernels performs better than the MERGILO baseline for all the considered metrics, showing a clear advantage of using the proposed frame similarities for reconciling knowledge graphs. The Wu-Palmer, Path and Leacock Chodorow measures use the inheritance relations only whereas Frame2Vec employs either graph walks or graph kernels over the FrameNet frame graph as well as subsumption hierarchy of FrameNet roles using either only FrameNet roles or improved subsumption hierarchy of FrameNet roles as introduced in Framester. Based on these settings, vector representations are generated which are further used for computing the cosine similarity. In general, Frame2Vec, for its intrinsic construction, exploits

[7] `http://wit.istc.cnr.it/stlab-tools/mergilo`

		muc	**bcub**	**ceafm**	**blanc**	**ceafe**
MERGILO Baseline		24.05	17.36	28.61	10.70	26.20
Similarity Measures						
Wu-Palmer		27.14	19.91	31.91	**12.81**	29.41
Path		27.16	19.93	31.85	12.73	29.38
Leacock Chodorow		27.04	19.80	31.74	12.77	29.21
Graph walks						
Frame2Vec	**Role2Vec**	**muc**	**bcub**	**ceafm**	**blanc**	**ceafe**
CBOW_200	CBOW_200	27.34	**19.99**	32.15	12.66	29.82
CBOW_200	SG_800	**27.38**	19.97	**32.29**	12.69	**29.98**
CBOW_200	SG_500	27.28	19.95	31.99	12.69	29.54
CBOW_200	CBOW_500	27.09	19.03	29.95	11.91	28.97
CBOW_500	SG_500	26.90	19.68	31.58	12.60	29.08
SG_200	SG_500	26.87	19.57	31.33	12.10	29.01
SG_500	SG_500	26.85	19.45	31.12	12.08	28.98
Graph kernels						
Frame2Vec	**Role2Vec**	**muc**	**bcub**	**ceafm**	**blanc**	**ceafe**
CBOW_200	CBOW_200	26.76	19.57	31.50	12.45	29.06
CBOW_200	CBOW_500	26.76	19.57	31.50	12.45	29.06
CBOW_200	SG_200	26.70	19.52	31.45	12.40	28.99
CBOW_200	SG_500	26.70	19.52	31.45	12.40	28.99
CBOW_500	CBOW_200	26.76	19.51	31.45	12.45	28.96
SG_200	CBOW_200	26.86	19.62	31.67	12.48	29.18
SG_500	CBOW_200	26.90	19.68	31.58	12.60	29.08

Table 1: Event-Based Knowledge Reconciliation Results. The best results are marked in bold.

more semantics than the other similarity measures (Wu-Palmer, Path and Leacock Chodorow); for such a reason, Frame2Vec provides the highest results for almost each evaluation measure except for BLANC.

BLANC is more sensitive to wrong assignments when clusters of mentions are larger, since a wrong assignment lead to a higher number of wrong non-coreference links. Therefore, although BLANC is case-by-case coherent with the other measures (when BLANC is low, the other measures are low and vice-versa), in the few cases when Frame2Vec is outperformed by other measures (Wu-Palmer, Path and Leacock Chodorow), the BLANC measure, and in particular the contribution given by non-coreference link, gives a much smaller score. These cases influence the overall average and for this reason in Table 1 BLANC seems to have a different behaviour than the other measures.

The generated models i.e., vector representations of FrameNet frames generated using FrameNet graph and subsumption hierarchy of FrameNet roles using RDF2Vec are freely available on-line[8].

[8] `http://lipn.univ-paris13.fr/~alam/Frame2Vec/`

6 Conclusions and Discussion

This paper presents a way to perform event-reconciliation for merging multiple event-oriented knowledge graphs originated from multiple texts. It uses existing tool MERGILO, a tool for reconciling knowledge graphs using word similarity and graph alignment. The current study exploits several path-based similarity measures for frames and semantic roles, i.e., following the approach *RDF2Vec*, graph-based frame embeddings were generated. The evaluation shows that the introduced approach is an effective improvement over the baseline.

Ongoing work concentrates on practical applications of frame embeddings in real systems, such as news series integration, knowledge graph evolution with robust event reconciliation (e.g. in streaming of texts where we expect relatedness or updates), or conflict detection across texts describing similar facts with different narratives or perspectives.

References

1. Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566. Citeseer, 1998.
2. Collin F Baker, Charles J Fillmore, and John B Lowe. The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90, 1998.
3. Cosmin Bejan and Sanda Harabagiu. Unsupervised event coreference resolution. *Comput. Linguist.*, 40(2):311–347, June 2014.
4. Alexander Budanitsky and Graeme Hirst. Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47, 2006.
5. Yun-Nung Chen, William Yang Wang, and Alexander I Rudnicky. Jointly modeling inter-slot relations by random walk on knowledge graphs for unsupervised spoken language understanding. In *HLT-NAACL*, pages 619–629, 2015.
6. Maddalen Lopez De Lacalle, Egoitz Laparra, and German Rigau. Predicate matrix: extending semlink through wordnet mappings. In *LREC*, pages 903–909, 2014.
7. Gerben Klaas Dirk de Vries and Steven de Rooij. Substructure counting graph kernels for machine learning from rdf data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:71–84, 2015.
8. Sourav Dutta and Gerhard Weikum. Cross-document co-reference resolution using sample-based clustering with knowledge enrichment. *Transactions of the Association of Computational Linguistics*, 3(1):15–28, 2015.
9. Jérôme Euzenat and Pavel Shvaiko. *Ontology matching*. Springer-Verlag, Heidelberg, 2nd edition, 2013.
10. Christiane Fellbaum, editor. *WordNet: an electronic lexical database*. MIT Press, 1998.
11. Aldo Gangemi, Mehwish Alam, Luigi Asprino, Valentina Presutti, and Diego Reforgiato Recupero. Framester: a wide coverage linguistic linked data hub. In *Knowledge Engineering and Knowledge Management: 20th International Conference, Bologna, Italy*, pages 239–254, 2016.

12. Aldo Gangemi, Valentina Presutti, Diego Reforgiato Recupero, Andrea Giovanni Nuzzolese, Francesco Draicchio, and Misael Mongiovi. Semantic Web Machine Reading with FRED. *Semantic Web Journal*, 2016.
13. Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. Sensembed: Learning sense embeddings for word and relational similarity. In *ACL (1)*, pages 95–105, 2015.
14. Karin Kipper Schuler. *Verbnet: A Broad-coverage, Comprehensive Verb Lexicon*. PhD thesis, Philadelphia, PA, USA, 2005. AAI3179808.
15. Simon Lacoste-Julien, Konstantina Palla, Alex Davies, Gjergji Kasneci, Thore Graepel, and Zoubin Ghahramani. Sigma: Simple greedy matching for aligning large knowledge bases. In *KDD2013*, pages 572–580, New York, USA, 2013. ACM.
16. Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500, 2012.
17. Xiaoqiang Luo. On coreference resolution performance metrics. In *Proc. of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 25–32. Association for Computational Linguistics, 2005.
18. Misael Mongiovì, Diego Reforgiato Recupero, Aldo Gangemi, Valentina Presutti, and Sergio Consoli. Merging open knowledge extracted from text with MERGILO. *Knowl.-Based Syst.*, 108:155–167, 2016.
19. Roberto Navigli and Simone Paolo Ponzetto. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193:217–250, 2012.
20. A. G. Nuzzolese, A. Gangemi, V. Presutti, P. Ciancarini, and A. Musetti. Automatic Typing of DBpedia Entities. In *Proc. of the International Semantic Web Conference (ISWC)*, Boston, MA, US, 2012.
21. Marta Recasens and Eduard Hovy. Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(04):485–510, 2011.
22. Diego Reforgiato Recupero. Efficient graph matching. *Encyclopedia of Data Warehousing and Mining*, pages 736–743, 2009.
23. Petar Ristoski and Heiko Paulheim. Rdf2vec: Rdf graph embeddings for data mining. In *ISWC*, pages 498–514, 2016.
24. Fabian M. Suchanek, Serge Abiteboul, and Pierre Senellart. PARIS: Probabilistic alignment of relations, instances, and schema. In *Proceedings of the VLDB Endowment*, volume 5, pages 157–168. VLDB Endowment, 2011.
25. Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Proceedings of the 6th conference on Message understanding*, pages 45–52, 1995.
26. Joshua T Vogelstein, John M Conroy, Vince Lyzinski, Louis J Podrazik, Steven G Kratzer, Eric T Harley, Donniell E Fishkind, R Jacob Vogelstein, and Carey E Priebe. Fast approximate quadratic programming for graph matching. *PLoS One*, 10(4):e0121002, 2015. PMID: 25886624.
27. William Yang Wang and Diyi Yang. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *EMNLP*, pages 2557–2563, 2015.