Relation- and Phrase-level Linking of FrameNet with Sar-graphs

Aleksandra Gabryszak, Sebastian Krause, Leonhard Hennig, Feiyu Xu, Hans Uszkoreit
LT Lab @ German Research Center for Artificial Intelligence
Alt-Moabit 91c, Berlin, Germany
{alga02,skrause,lehe02,feiyu,uszkoreit}@dfki.de

Abstract

Recent research shows the importance of linking linguistic knowledge resources for the creation of large-scale linguistic data. We describe our approach for combining two English resources, *FrameNet* and *sar-graphs*, and illustrate the benefits of the linked data in a relation extraction setting. While FrameNet consists of schematic representations of situations, linked to lexemes and their valency patterns, sar-graphs are knowledge resources that connect semantic relations from factual knowledge graphs to the linguistic phrases used to express instances of these relations. We analyze the conceptual similarities and differences of both resources and propose to link sar-graphs and FrameNet on the levels of relations/frames as well as phrases. The former alignment involves a manual ontology mapping step, which allows us to extend sar-graphs with new phrase patterns from FrameNet. The phrase-level linking, on the other hand, is fully automatic. We investigate the quality of the automatically constructed links and identify two main classes of errors.

Keywords: Linking linguistic resources, knowledge graphs, relation extraction

1. Introduction

Linguistic linked open data (Chiarcos et al., 2013) is the idea and movement of publishing linguistic resources according to the *linked data principles* (Bizer et al., 2009). A prerequisite for releasing such data is to identify semantically corresponding elements in distinct datasets. We describe our work on linking two English linguistic resources, one from the area of lexical semantics (*FrameNet*) and one being a repository of linguistic expressions for knowledge-graph relations (*sar-graphs*)[1].

A *sar-graph* is a graph containing linguistic knowledge at syntactic and lexical semantic levels for a given language and target relation. Sar-graphs (Uszkoreit and Xu, 2013) are a semi-automatically created resource which explicitly links the semantic relations of knowledge graphs (Dong et al., 2014; Lehmann et al., 2015) to the linguistic patterns which can express these relations in natural-language text. The current version of sar-graphs contains syntactic dependency relations between content words, word senses, and semantic arguments. The linguistic patterns in sar-graphs are automatically acquired via a pattern discovery method based on distant supervision (Mintz et al., 2009; Krause et al., 2012). Thus sar-graphs can be directly applied to free texts for relation extraction.

Early work on lexical-semantics resources has focused on gathering information about individual words and their different meanings in varying contexts, the famous example being WordNet (Fellbaum, 1998). Linguistic knowledge resources that go beyond the level of lexical items are scarce and of limited coverage due to significant investment of human effort and expertise required for their construction. FrameNet (Baker et al., 1998) is such a resource and provides fine-grained semantic relations of predicates and their arguments. However, FrameNet does not provide an explicit link to real-world fact types.

There is increasing research in automatically creating large-scale linguistic resources, often these have been built on top of existing resources. For example, BabelNet (Navigli and Ponzetto, 2012) merged Wikipedia concepts including entities with word senses from WordNet; a similar strategy was pursued in ConceptNet (Speer and Havasi, 2013). Only few approaches have included FrameNet in their linking efforts (Scheffczyk et al., 2006; Bonial et al., 2013; Aguilar et al., 2014). A particular example is UBY (Gurevych et al., 2012), which provides a standardized representation for several combined lexico-semantic resources via the Lexical Markup Framework. None of these approaches linked FrameNet both to knowledge-graph relations and extended it with linguistic patterns at the same time.

Much of the recent literature has dealt with the problems of semantic role labeling and frame-semantic parsing (Gildea and Jurafsky, 2002; Das et al., 2014; FitzGerald et al., 2015), i.e., the automatic enrichment of sentences with FrameNet-style annotation. Often, these systems suffer from a lack of training data. Although several ideas to address this issue have been presented (Giuglea and Moschitti, 2006; Pavlick et al., 2015; Chang et al., 2015), the problem largely remains unsolved. Our approach can support these systems by increasing the amount of available training data.

In previous work (Krause et al., 2015), we have linked sar-graphs to word-level lexical-semantic resources like BabelNet. We continue this line of work and describe in this paper the ongoing effort of linking the data-driven sar-graphs with the curated FrameNet. We also show that by enriching sar-graphs with FrameNet data we can mitigate the notorious long-tail distribution of linguistic phrases, which allows us to reach higher coverage in extraction experiments.

In the following, we discuss two ways of linking sar-graphs with FrameNet, which are in spirit of the large-scale efforts mentioned above. We believe that both resources and their respective applications can benefit from the coupling. Our contributions are as follows:

- We present a *phrase-level linking* of FrameNet and sar-graphs, which automatically identifies corresponding sentence templates which indicate semantics shared by FrameNet frames and sar-graph patterns, thereby

[1] http://sargraph.dfki.de

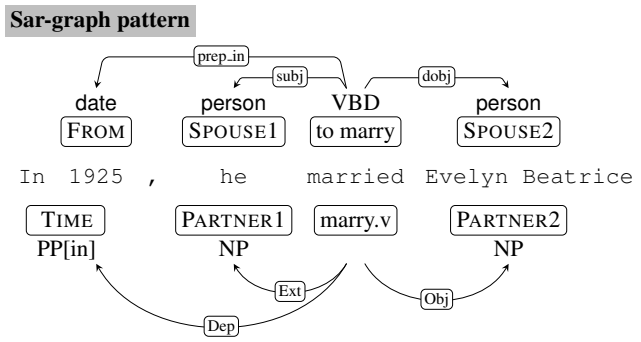

Figure 1: Comparison of pattern formalisms from sar-graphs (top) and FrameNet (bottom). Both representations connect semantic arguments (FROM, SPOUSE1, SPOUSE2, TIME, PARTNER1, PARTNER2) and lexical items ("to marry", "marry.v") via grammatical relations ("prep_in", "subj", "dobj", "Dep", "Ext", "Obj").

tightly coupling both resources.

- We describe our evaluation methodology for the phrase-level linking and report first results.
- Furthermore, we discuss our strategy for *frame-relation linking*, which involves a manual ontology mapping step.
- We illustrate the usefulness of this second approach to linking by applying patterns learned from the FrameNet data for relation extraction, one of the core applications of sar-graphs.

2. The two resources

Sar-graphs Sar-graphs (DFKI LT Lab, 2015) are directed multi-graphs. They are specific to a particular language and knowledge-base relation (e.g., the *marriage* relationship). The linguistic constructions contained in sar-graphs are modeled as sub-trees of dependency-graph representations of sentences. Each structure typically describes one particular way to express the relation. The graphs are created in a data-driven way by mining relation mentions from the web and discovering patterns from the dependency structures of the sentential mentions, applying an automatic filtering step (Moro et al., 2013) for high-confidence patterns, and finally superimposing and partially merging the relation paraphrases. The top of Figure 1 depicts an individual pattern from the *marriage* sar-graph, in which the semantic arguments of the target relation are labelled with grammatical functions in a dependency tree. In reality, sar-graphs consist of thousands of such patterns. Currently, sar-graphs are available in English for 25 semantic relations.

FrameNet The FrameNet Project (The Berkeley FrameNet project, 2010) has created a lexical resource for English that documents the range of semantic and syntactic combinatorial possibilities of words and their senses. FrameNet consists of schematic representations of situations (called frames), e.g., the frame *win prize* describes an awarding situation with semantic roles (frame elements, FE), like COMPETITOR, PRIZE, COMPETITION, etc. A pair of word and frame forms a lexical unit (LU), similar to a word sense in a thesaurus. LUs are connected to lexical entries (LEs), which capture the valence patterns of frames, providing information about FEs and their phrase types and grammatical functions in relation to the LUs. Each pattern is illustrated by a set of annotated sentences. An example valence pattern is shown in the bottom of Figure 1.

Comparison of FrameNet to sar-graphs Sar-graphs resemble frames in many aspects, e.g., both define semantic roles for target concepts and provide detailed valency information for linguistic constructions referring to the concept. However, there are some differences. FrameNet contains a number of very generic frames (e.g., *forming_relationships*) that have no explicit equivalent in a sar-graph relation. The database-driven sar-graphs also specify fewer semantic roles than frames typically do, covering mainly the most important aspects of a relational concept from a knowledge-base population perspective. For example, the sar-graph for *marriage* lists arguments for the SPOUSEs, LOCATION and DATE of the wedding ceremony as well as a DIVORCEDATE, while the related frame *forming_relationships* additionally covers, e.g., an EXPLANATION (divorce reason, etc.) and an ITERATION counter (for the relationships of a person).
Above that, FrameNet specifies relations between frames (*inheritance*, *subframe*, *perspective on*, *using*, *causative of*, *inchoative of*, *see also*) and connects in this way also the lexical units evoking the related frames. For example, frames *commerce_buy* and *commerce_sell* represent perspectives on the frame *commerce_good_transfer*, and link by the same relation the verbs `to sell` and `to buy`. Sar-graphs are currently not linked to one another.
Another difference is the relationship between lexical items and their corresponding frames/sar-graph relations. LUs in FrameNet imply frames by subsumption, e.g., `to befriend` and `to divorce` are subsumed by *forming_relationships*. In comparison, sar-graphs cluster both expressions that directly refer to instances of the target relation (e.g., `to wed` for *marriage*) and those that only entail them (e.g., `to divorce` for *marriage*). This entailment is, in turn, partly represented in FrameNet via frame-to-frame relations like *inheritance*, *cause* and *perspective*.
We presented a more detailed comparison of the resources in (Krause et al., 2015).

Linking sar-graphs to FrameNet Based on the similarities between FrameNet and sar-graphs we propose to link the resources on phrase level (Section 3.) and on frame-relation level (Section 4.).

3. Phrase-level linking

Here, we describe our methodology for linking FrameNet and sar-graphs on the level of phrases.
FrameNet 1.5 contains 74k valency patterns and more than 170k annotated sentences. We link them to two variants of the sar-graphs, an automatically filtered version (~300k phrase patterns) and a curated subset (~4.2k). Instead of directly aligning the valency patterns with the corresponding dependency patterns, we apply the sar-graph pattern discovery pipeline to the FrameNet sentences associated with the

valency patterns. The phrase patterns extracted from the FrameNet sentences are then matched with the sar-graph patterns and serve as a proxy for the linking. Thus, we can avoid the painful mapping of the two syntax representations.

Approach We filter the set of annotated sentences in FrameNet for those that mention two or more frame elements. These sentences are then processed by a dependency parser, after which sar-graph-like phrase patterns are extracted. We extracted more than 80k *FrameNet patterns*.
We determine corresponding patterns from sar-graphs and FrameNet by comparing them via tree edit distance.[2] We only take into account the lexical level and syntax of the patterns and ignore differences in the definition of semantic arguments and their names, as these would be hard to resolve automatically and would constitute an ontology integration step (see Section 4.). The calculated distance d between two patterns is normalized by the number of edges in the two patterns. This allows us to order the links between FrameNet patterns and sar-graph patterns by d and to discard all links with $d >$ threshold t. We furthermore exclude all links where either of the patterns does not mention the source lexical unit of the original valence pattern.
We define three classes of pattern-level links with examples in Table 1:

- *Exact match*: $d = 0.0$. The link is correct if the patterns are semantically equivalent.
- *Subsumption*: $d > 0.0$ and one of the two patterns is syntactically fully contained in the other. Correctness of the link requires that there is an entailment relation between the patterns.
- *Other*: $d > 0.0$ and neither pattern is included in the other. The link is correct if the meanings of the patterns are related.

Statistics and linking errors We have conducted the linking step and present the distribution of links across the three classes in Table 2, for a threshold of 0.5 on the normalized distance d. A large fraction of sar-graph elements were automatically aligned with their FrameNet counterparts. Since the linking step is currently based solely on the lexical and syntactic features of the patterns, there are two main causes for semantically erroneous links:

- *Semantic ambiguity*: Linked patterns are not synonymous due to polysemy/homonymy. For example, a sar-graph pattern for relation *organization leadership* which contained the lemma `to lead` was erroneously linked to a pattern from frame *cotheme*, which uses this verb to mean *showing someone the way* and not, as in *organization leadership*, *to be in charge of something*.
- *Argument-type mismatch*: Patterns have a different meaning because of the semantic types of their respective arguments. For example, the ORGANIZATION of sar-graph patterns for *organization leadership* was matched to the element DEPICTIVE of frame *leadership*, where the correct mapping would have been to frame element GOVERNED. Consider the difference between `Informatica chairman Sohaib Abbasi` and `deputy chairman Eric Goodman`.

[2] We used the algorithm by Zhang and Shasha (1989) as provided at `http://web.science.mq.edu.au/~swan/howtos/treedistance/package.html`.

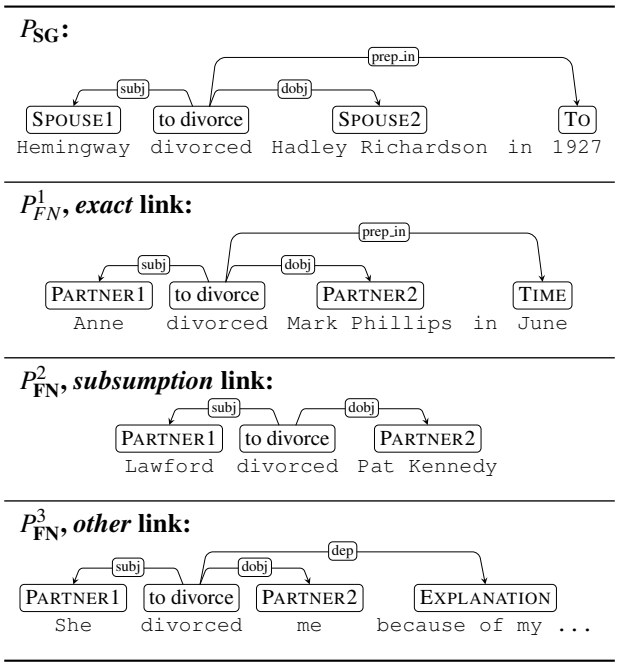

Table 1: Pattern-link examples. P_{SG} is a sar-graph pattern for relation *marriage*; P^1_{FN}, P^2_{FN}, P^3_{FN} are FrameNet patterns from *forming relationships* that were linked to P_{SG}.

Sar-graph variant	Exact	Subsumption	Other
Curated	251	554	4,419
Autom. filtered	2,978	8,329	113,201

Table 2: Distribution of pattern links.

4. Linking frames and relations

We now discuss how we integrate parts of the schemas underlying the two resources, i.e., how we identify a subset of the 1,019 frames in FrameNet 1.5 which correspond to the 25 semantic relations for which sar-graphs are available.

Approach The ontology-mapping is conducted manually as follows: for each of the sar-graphs we determine which frames have a similar meaning by comparing their respective definitions and aligning the arguments of sar-graphs with the frame elements. We focus on mapping the *essential arguments* from sar-graphs (e.g., ORGANIZATION and PERSON in the *employment tenure* relation) on *elements* from FrameNet (e.g., EMPLOYEE and EMPLOYER for frame *employment start*).
The mapping of frames to relations is a many-to-many mapping, e.g., the relation *employment tenure* is mapped to 22 frames, among them the frame *leadership*. This frame is in turn linked to the sar-graph relations *organization leadership* and *organization membership*. Figure 2 shows an excerpt from the frame-relation alignment.

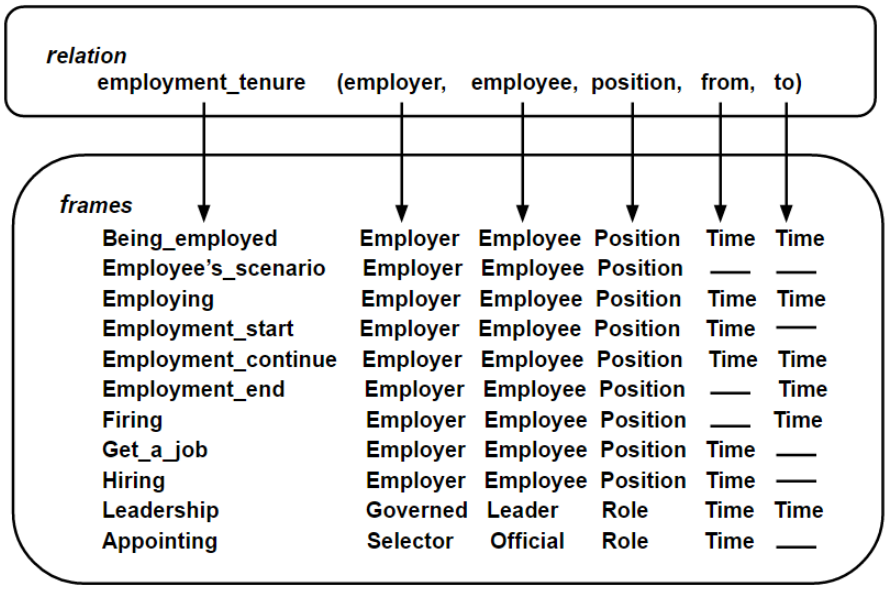

Figure 2: Linked frames for sar-graph relation *employment tenure*.

Sar-graph variant	FrameNet phrases	# Correct extractions	Recall improvement	Precision loss
Curated	no	42,639	—	—
Curated	yes	67,680	+58.73%	-37.64%
Autom. filtered	no	174,063	—	—
Autom. filtered	yes	184,343	+5.91%	-9.18%

Table 3: Results from extraction experiment on ClueWeb.

We mapped 25 sar-graph relations to 260 frames, with the number of frames per sar-graph ranging from 1 to 40. Some of the more extreme cases are relation *siblings*, which is linked only to the frame *kinship*, and relation *acquisition*, which is mapped to lexical frames like *commerce buy*, *commerce sell*, *shopping*, *receiving*, *getting*, *possession*.

The semantic agreement and mutual coverage of an identified pair of frame and relation varies greatly. *Acquisition* has a largely congruent extent with frames *commerce buy*, *commerce sell*, and *shopping*. In contrast, the frame *getting* is more general than *acquisition*, e.g., does not require payment for acquired entities and also contains patterns not at all related to transaction of physical goods. However, *getting* also contains lexical units like `to acquire` and `to get`, which can be useful in contexts implying commercial business and buyer and seller roles, e.g., `Yahoo acquired Polyvore` or `Peter got the novel from Amazon`.

Extraction experiment We evaluated the impact of expanding sar-graphs with FrameNet phrases with a relation extraction experiment. In particular, we were interested in whether the addition would substantially increase the coverage of linguistic expressions. We selected a set of approx. 30 million sentences from the ClueWeb datasets[3] (The Lemur Project, 2012) with linked mentions of Freebase entities (Gabrilovich et al., 2013).

All sar-graph patterns were matched against the sentences of the corpus in order to extract facts, as were the FrameNet phrases which were part of a frame linked to a sar-graph relation. We evaluated the detected relation mentions by checking whether they were listed in Freebase. Table 3 displays the amount of correct facts that the two used variants of sar-graphs covered, as well as the amount of them extracted after the addition of FrameNet phrases. We can see that for both sar-graph variants, the extraction performance substantially improves after the expansion step.

5. Conclusion and outlook

In this paper we presented the current state of our ongoing work on linking our data-driven resource of linguistic expressions for knowledge-base relations with FrameNet on two different levels of abstraction, i.e., on the phrase level and on the level of frames/relations.

Regarding the phrase-level linking, we described our evaluation methodology for the automatic aspect of the linking process; in the immediate future we will use this to estimate the quality of the established links. In the medium term, we would like to combine this line of work with approaches for relation-taxonomy induction from sar-graphs, which could help to automatically refine the hierarchy of frames in FrameNet.

Furthermore, we showed that the manual relation-phrase level linking can feed new relation paraphrases to the sar-graphs, which boosts their performance in tasks like relation extraction.

[3] `http://www.lemurproject.org/`

6. Acknowledgements

This research was partially supported by the German Federal Ministry of Education and Research (BMBF) through the project ALL SIDES (01IW14002) and by Google through a Focused Research Award granted in July 2013.

7. Bibliographical References

Aguilar, J., Beller, C., McNamee, P., Van Durme, B., Strassel, S., Song, Z., and Ellis, J. (2014). A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards. In *Proceedings of the Second Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 45–53. Association for Computational Linguistics.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). 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. Association for Computational Linguistics.

Bizer, C., Heath, T., and Berners-Lee, T. (2009). Linked data - The story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22.

Bonial, C., Stowe, K., and Palmer, M. (2013). Renewing and revising SemLink. In *Proceedings of the 2nd Workshop on Linked Data in Linguistics (LDL-2013): Representing and linking lexicons, terminologies and other language data*, pages 9 – 17. Association for Computational Linguistics.

Chang, N., Paritosh, P., Huynh, D., and Baker, C. (2015). Scaling semantic frame annotation. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 1–10. Association for Computational Linguistics.

Chiarcos, C., McCrae, J., Cimiano, P., and Fellbaum, C. (2013). Towards open data for linguistics: Linguistic linked data. In Alessandro Oltramari, et al., editors, *New Trends of Research in Ontologies and Lexical Resources*, Theory and Applications of Natural Language Processing, pages 7–25. Springer Berlin Heidelberg.

Das, D., Chen, D., Martins, A. F. T., Schneider, N., and Smith, N. A. (2014). Frame-semantic parsing. *Computational Linguistics*, 40(1):9–56.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., and Zhang, W. (2014). Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*, pages 601–610. ACM.

Christiane Fellbaum, editor. (1998). *WordNet: an electronic lexical database*. Christiane Fellbaum.

FitzGerald, N., Täckström, O., Ganchev, K., and Das, D. (2015). Semantic role labeling with neural network factors. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 960–970, Lisbon, Portugal. Association for Computational Linguistics.

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Comput. Linguist.*, 28(3):245–288, 9.

Giuglea, A.-M. and Moschitti, A. (2006). Semantic role labeling via FrameNet, VerbNet and PropBank. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pages 929–936. Association for Computational Linguistics.

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). UBY - A large-scale unified lexical-semantic resource based on LMF. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590. Association for Computational Linguistics.

Krause, S., Li, H., Uszkoreit, H., and Xu, F. (2012). Large-scale learning of relation-extraction rules with distant supervision from the web. In *Proc. of 11th ISWC, Part I*, pages 263–278.

Krause, S., Hennig, L., Gabryszak, A., Xu, F., and Uszkoreit, H. (2015). Sar-graphs: A linked linguistic knowledge resource connecting facts with language. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 30–38. Association for Computational Linguistics.

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Hellmann, S., Morsey, M., van Kleef, P., Auer, S., and Bizer, C. (2015). DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Moro, A., Li, H., Krause, S., Xu, F., Navigli, R., and Uszkoreit, H. (2013). Semantic rule filtering for web-scale relation extraction. In *Proceedings of the 12th International Semantic Web Conference, ISWC 2013, Part I*, volume 8218 of *Lecture Notes in Computer Science*, pages 347–362. Springer.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Pavlick, E., Wolfe, T., Rastogi, P., Callison-Burch, C., Dredze, M., and Van Durme, B. (2015). FrameNet+: Fast paraphrastic tripling of FrameNet. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 408–413. Association for Computational Linguistics.

Scheffczyk, J., Pease, A., and Ellsworth, M. (2006). Linking FrameNet to the Suggested Upper Merged Ontology. In *Formal Ontology in Information Systems, Proceedings of the Fourth International Conference, FOIS 2006*, volume 150 of *Frontiers in Artificial Intelligence and Applications*, pages 289–300. IOS Press.

Speer, R. and Havasi, C. (2013). ConceptNet 5: A large semantic network for relational knowledge. In Iryna Gurevych et al., editors, *The People's Web Meets NLP*, Theory and Applications of Natural Language Processing, pages 161–176. Springer Berlin Heidelberg.

Uszkoreit, H. and Xu, F. (2013). From strings to things - Sar-graphs: A new type of resource for connecting knowledge and language. In *Proceedings of the NLP & DBpedia workshop co-located with the 12th International Semantic Web Conference (ISWC 2013)*, volume 1064 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Zhang, K. and Shasha, D. (1989). Simple fast algorithms for the editing distance between trees and related problems. *SIAM J. Comput.*, 18(6):1245–1262, 12.

8. Language Resource References

DFKI LT Lab. (2015). *Sar-graphs: Graphs of Semantically Associated Relations*. German Research Center for Artificial Intelligence, version 3.0.

Evgeniy Gabrilovich and Michael Ringgaard and Amarnag Subramanya. (2013). *FACC1: Freebase annotation of ClueWeb corpora*. Version 1 (Release date 2013-06-26, Format version 1, Correction level 0).

The Berkeley FrameNet project. (2010). *FrameNet*. International Computer Science Institute, Berkeley, California, version 1.5.

The Lemur Project. (2012). *ClueWeb*. Center for Intelligent Information Retrieval at the University of Massachusetts, Amherst, and the Language Technologies Institute at Carnegie Mellon University, versions 09 & 12.