A Regional News Corpora for Contextualized Entity Discovery and Linking

Adrian M.P. Braşoveanu*, Lyndon J.B. Nixon*, Albert Weichselbraun†, Arno Scharl*
* Department of New Media Technology, MODUL University Vienna,
Am Kahlenberg 1, 1190 Vienna, Austria
† Swiss Institute for Information Research, University of Applied Sciences Chur,
Pulvermühlestrasse 57, 7000 Chur, Switzerland
{adrian.brasoveanu,lyndon.nixon,arno.scharl}@modul.ac.at
albert.weichselbraun@htwchur.ch

Abstract

This paper presents a German corpus for Named Entity Linking (NEL) and Knowledge Base Population (KBP) tasks. We describe the annotation guideline, the annotation process, NIL clustering techniques and conversion to popular NEL formats such as NIF and TAC that have been used to construct this corpus based on news transcripts from the German regional broadcaster RBB (Rundfunk Berlin Brandenburg). Since creating such language resources requires significant effort, the paper also discusses how to derive additional evaluation resources for tasks like named entity contextualization or ontology enrichment by exploiting the links between named entities from the annotated corpus. The paper concludes with an evaluation that shows how several well-known NEL tools perform on the corpus, a discussion of the evaluation results, and with suggestions on how to keep evaluation corpora and datasets up to date.

Keywords: Named entity linking, corpora, contextualization, knowledge base population, television news transcripts, evaluation.

1 Introduction

With an ever increasing amount of news and social media coverage comes a need to extract relevant information automatically. A method that can help in this endeavour is Named Entity Linking (NEL), the grounding to knowledge base entries of the mentions extracted from texts. Knowledge Base Population (KBP) (Ji et al., 2014), the process through which the entities identified by NEL systems are used to populate new knowledge bases, is another useful technique for exploring the relations between entities. However, since NEL or KBP evaluation tasks might require a new corpus or at least a new gold standard, and the creation of such resources requires significant effort, there is a desire to automate steps in the corpus creation process. By exploiting the knowledge graph built with named entities from a corpus, new gold standards can be created for specific tasks like KBP, entity contextualization or enrichment. In contrast to research focused on social media (Bontcheva and Rout, 2012), we used regional news and analyzed transcripts from the German regional broadcaster RBB (Rundfunk Berlin Brandenburg). This paper presents the corpus designed for KBP evaluations and shows how new datasets for specific evaluation tasks are automatically created by exploiting the links between the entities from the graph associated with the corpus.

The rest of this paper is organized as follows: Section 2 presents the dataset and the process used to generate it: creating annotations; calculating inter-annotator agreements; NIL clustering of the unlinked entities; and conversion to popular NEL evaluation formats such as TAC KBP (Ji et al., 2014) and NIF (Hellmann et al., 2013). Section 3 describes the knowledge engineering methods used for generating new datasets from the entity graph built from the corpus for tasks like ontology population and enrichment, contextualized entity extraction or improving graph disambiguation. Section 4 evaluates the performance of multiple NEL tools on the presented corpus. The article concludes with an analysis of corpus refinement strategies deployed during the creation of the corpus, and with a discussion of how to apply the techniques presented in this article to the automated generation of large datasets and corpora for Natural Language Processing (NLP).

2 Corpus Construction

The European research project LinkedTV[1] has collected a set of high-quality transcripts of German news broadcasts from the regional broadcaster RBB (Rundfunk Berlin Brandenburg) with the purpose of evaluating NEL with the *Recognyze* component (Weichselbraun et al., 2015). The RBB corpus contains subtitles extracted from the RBB news show *Abendschau* (daily news broadcast between 19:30 and 20:00 CET).[2] While there are several other German NEL corpora, this corpus being extracted from local news contains localized geographical information (street names, neighborhoods, highways, etc), person names that are not necessarily famous enough to be included in Wikipedia or in the large Knowledge Bases, local branches of national or international organizations, and events that are important to the local community. Also abbreviations tend to be used more often in television content than in the news media (news articles, blog posts, etc).

Document Selection. The RBB transcripts represent a set of local news items from Berlin and Brandenburg, and are focused on a wide array of news topics such as floods, traffic jams, immigration, sports and political events, and local administration. Due to the regionality of the content, frequent use of shortened names for entities, direct or indirect references to local or historical events

[1] http://linkedtv.eu
[2] https://github.com/linkedtv/videocorpus

(e.g. anniversaries of the 1953 East German Uprising or of Kennedy's visit to Berlin from 1963) and due to the differences between the written German from newspapers or blogs and the German spoken in TV shows which is closer to the German spoken in the real world, the RBB content presents some interesting challenges for NEL tasks.

Annotation Rules similar to the ones from TAC KBP and ACE for annotating the corpora (Nuzzolese et al., 2015) were used. The ontology used by the annotators contained the following main classes: *Person*, *Organization*, *Location*, *Event*, *Work*, *Product*, *Miscellaneous*. For each class we have presented suggestions for subclasses, a set of rules and minimal guidelines for the annotators. The human annotators were asked to take the subclass suggestions more as a guideline and they were free to consider any other subtype for inclusion in the main entity types. Since Person, Organisation and Location classes are well-known within the community, we provide here some suggestions for subclasses and examples from the rules we used for the second set of classes (Event, Product, Work):

- **Usual Subclasses for Events** include attack, election, protest, military conflict, scandal, sports event, terrorist attack, war, national holidays, concerts, launches.

  Different Names for the Same Event - All the surface forms for an event should be marked with the corresponding entity (e.g., Battle of Damascus, Operation Damascus Volcano).

- **Usual Subclasses for Products** include engine, airplane, car, ship, spacecraft, train, camera, phone, computer, software, game, instrument, weapon, magazine, newspaper, social network, food, brand.

  Product and Company Name - If the name of a company and its product are identical (e.g., Google, Facebook), annotate according to the context - e.g., Facebook product if the text mentions social network.

- **Usual Subclasses for Works (of Art)** include film, play, TV show (including TV series), written work, music, entertainment, sculpture, painting, book, game.

  Mention of Authorship - Mentions of classic works can be accompanied by the name of their creator (e.g., Da Vinci's Mona Lisa, Ravel's Bolero). If this happens, include author's name in the annotation.

All the entity mentions that were not part of the previously mentioned classes were marked as Misc (e.g., species of plants or animals, temporal types, diplomas, body parts, genes, units of measurements, etc.).

Annotation Process. The subtitles were imported into GATE (Cunningham et al., 2011) and manually annotated with these main classes by several annotators using the ontology. The annotators provided information on surface forms, entity types and German DBpedia links (Hellmann et al., 2012) if they were available and were asked to comment on all the encountered problems. An expert in Linked Data (Berners-Lee et al., 2009) and NEL worked in close contact with the annotators to judge the problematic cases (e.g., different surface forms, missing DBpedia links).

Feature	**Count**
Number of clips	150
Avg clip duration (seconds)	132.45
Avg no of entity mentions per document	12.36
Avg no of DBpedia linked entities per document	8.97
Persons	313
Organisations	254
Geo-Political Entities	733
Events	116
Products	13
Works	38
Misc	387

Table 1: Basic statistics of the corpus

Entity Type	**Clusters**
Persons	70
Organisations	41
Geo-Political Entities	52

Table 2: NIL clustering statistics

Agreement. After the problematic cases were solved, the quality of the agreement for entity typing and linking was calculated. While there can be arguments against the use of German DBpedia, we consider that texts in a specific language should be annotated in the same language except if specified otherwise. We have used the same knowledge base version (German DBpedia 2015) during all the stages of the corpus creation and evaluation. Additional evaluation tasks can later be added to search for the unlinked entities in other versions or languages of the same or different knowledge bases.

Export. Several modules help us convert the results to different modern formats like TAC KBP and NIF. For TAC KBP style evaluations (Ji et al., 2014), the NIF version is restricted to types contained in such evaluations (Person, Organization, Geo-Political Entity). The datasets described in Section 3 and the evaluation of Section 4 are built on top of this corpus version. Depending on the evaluation and the tools used for the experiments, various versions of the corpus can be provided: *full* (all entity types), *KBP* (restricted to only the three main types: Person, Location, Organisation), *NIF or CSV with or without NIL entities* - without NIL entities version is especially useful for GERBIL evaluations (Usbeck et al., 2015).

NIL Clustering. Entities that were not linked by the annotators were marked as NIL and clustered based on their types and mention strings with the following algorithms: naive,

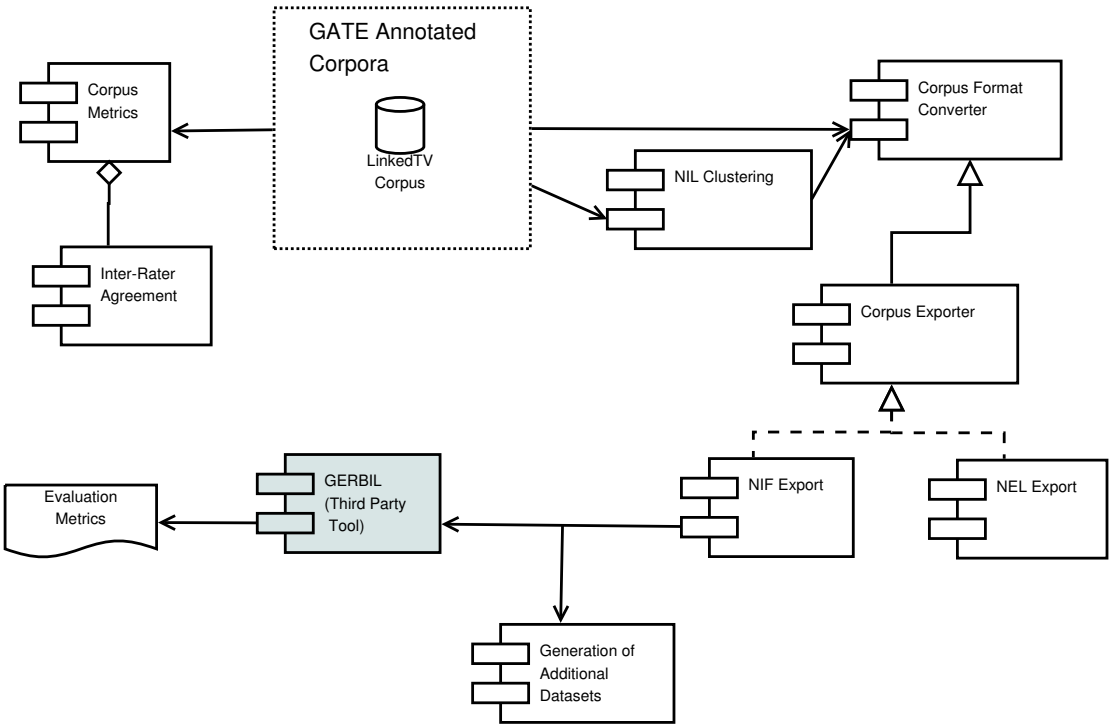

Figure 1: Framework used for corpus construction

hierarchical, co-reference head. This is not a trivial task, as many entities might not be linked because of spelling mistakes, obscurity or the complicated disambiguation process. The naive algorithm considers that only mentions with the exact same string belong to the same clusters, the hierarchical clustering uses the Levenshtein distance to calculate the similarity between strings and groups them accordingly, while the co-reference algorithm improves a bit on the hierarchical algorithm and restricts the entities that belong to a cluster to the mentions that share the same head. We found that both hierarchical and co-reference head algorithms performed well on our dataset. The clusters were manually checked by the Linked Data expert. The entities that were not assigned to the right clusters were then extracted to their own clusters. If the number of features considered increases, NIL clustering can also be used for creating new entities by specifying some of their properties or relations from the data at hand. These extracted entities can later be linked by querying the knowledge base for the properties that were already extracted. Every time we change the underlining datasets used in evaluation (e.g., when we switch from DBpedia 2015 to DBpedia 2016) it is also recommended to run such queries in order to discover the newly added entities.

Basic Statistics. Some basic statistics (averages, counts) about the RBB corpus are presented in Table 1. All corpus documents contain at least one entity, but not all the entities are necessarily linked as the German DBpedia contains a much smaller set of entities than the English DBpedia. The number of NIL clusters for the three main entity types required in TAC KBP evaluations are presented in Table 2. Each cluster corresponds to an entity. Many people could not be found in DBpedia, as regional news often cover local individuals who might not be famous enough to warrant the inclusion in a knowledge base. The small number of unlinked geo-political entities suggests that DBpedia has good coverage of these entities in German.

3 Knowledge Engineering

Since the creation of corpora and datasets tends to be a costly process, we advocate the use of knowledge engineering methods to create new datasets and evaluations, as in time this leads to economies of time and budget.

Ontology and Knowledge Base Enrichment. The automated creation of subtypes for the main types of the ontology is one of the most useful methods for enriching both the ontology and the Knowledge Base we plan to use. The annotation guideline provided the annotators with some hints regarding the subtypes that might appear for each main type, but the ontology itself only contained the primary types. In order to create the subtypes we import the corresponding candidate DBpedia subtypes (rdf:type) from the entities identified in the texts. Main candidate types can also be added by extracting the typing of the entities marked as Misc. By using a threshold, the less used subtypes can be easily pruned, therefore, assuring that each ontology subtype will be populated.

Exploiting Links and Context for Graph Disambiguation. Modern approaches towards NEL rely on algorithms that exploit the mention-entity graphs constructed from the texts (Hoffart et al., 2011) and aim to identify a single correct disambiguation for each mention-entity pair. To create a gold standard for such graph disambiguation methods, we use the entity graph associated with the

corpus documents, contextual information (list of entities that co-occur with an entity, for example) and DBpedia PageRanks to select the most likely links between two entities. It has to be noted that context can be defined in multiple ways: as a similarity measure used to compare local context (words from the text) to Wikipedia categories (Bunescu and Pasca, 2006); as all the words from the text fragment (Hoffart et al., 2011); or even as the main topic of the text fragment (*sports* or *politics* (Cattoni et al., 2012), for example). While each method has its advantages, for the purpose of this article we will consider context as being the main topic of the text fragment together with the links between the identified entities. If context is defined like this, it can later be used to create evaluations that test the correct disambiguation path for the named entities. The output of the method used for automatically creating graph disambiguation paths is a list of connected surface forms, and if needed also the main topic that unites them (e.g., Apple and Steve Jobs when the main topic is computer).

Relation Extraction. It is straightforward to automatically extract links between entities and create an automated gold standard for relation extraction starting from any annotated corpora. The only serious issue is that that not all the extracted links are necessarily useful. A link like rdfs:seeAlso only tells us that there is a connection between two entities, but does not offer any particular detail about this connection regardless of the context in which it appears, whereas an owl:differentFrom link in a context where two entities have similar names points to the fact that these two entities are indeed different. To automatically curate the extracted relations and create the relation extraction gold standard, the types of links that do not necessarily offer important information about entities need to be blacklisted. The output of the method is a set of triples that represent the extracted relations. If required, one can always add more information about the entities or their surface forms and turn this into a gold standard for KBP as well. Such a gold standard is important for evaluating graph disambiguation tools like AIDA (Hoffart et al., 2011) or Babelfy (Navigli and Moro, 2014).

Figure 2 presents the relations among entities identified in a text and the expected output of the described knowledge engineering methods in a graph disambiguation context.

Actionable Knowledge. After creating corpora and using them to compare named entity linking components, future work will focus on improving the algorithms used for the disambiguation process. Besides testing the tool again when these algorithms are available, a good method to convey actionable knowledge is to analyze the documents as part of a visual dashboard. Recognyze has been integrated into the webLyzard Web intelligence platform (Scharl et al., 2016b; Scharl et al., 2016a), which uses a set of visualization components to explore the connections between documents, entities and events.

Type	Tool	P	R	F1
Person	Babelfy	0.61	0.40	0.48
	Spotlight	0.25	0.35	0.29
	Dexter	0.19	0.18	0.19
	Recognyze	**0.64**	**0.40**	**0.49**
Organisation	Babelfy	0.43	0.39	0.23
	Spotlight	**0.32**	**0.29**	**0.30**
	Dexter	0.16	0.07	0.10
	Recognyze	0.26	0.16	0.20
Location	Babelfy	0.45	0.24	0.31
	Spotlight	**0.31**	**0.42**	**0.36**
	Dexter	0.22	0.22	0.2

Table 3: Evaluation of entity linking performance

4 Experiments

Corpus evaluation. The evaluation was performed without taking into consideration NIL entities and draws upon the following four named entity linking systems: Babelfy (Navigli and Moro, 2014), Spotlight (Daiber et al., 2013), Dexter (Ceccarelli et al., 2013), and Recognyze (Weichselbraun et al., 2015). GERBIL (Usbeck et al., 2015) includes more annotation services (annotators), but it does not provide typed evaluations. A choice was made to select only those annotators that return German results through their online endpoints or whose results had good conversion scores from English to German DBpedia.

DBpedia Spotlight (Daiber et al., 2013) is well-known within the Semantic Web and NLP communities for being one of the first tools to use DBpedia and offer semantic approaches to the named entity recognition and disambiguation problems. It was built around a vector space model and is available through a public endpoint [3].

Babelfy (Navigli and Moro, 2014) was one of the first graph disambiguation tools that worked in a multilingual setting and it was built around the idea of word sense disambiguation. It offers a free webservice [4] with a limited number of requests and the possibility to evaluate it for research purposes.

Dexter (Ceccarelli et al., 2013) is an entity disambiguation framework that was built in order to simplify NEL approaches. It is ideal as a basis for evaluating NEL algorithms and also available as a free webservice [5].

Recognyze (Weichselbraun et al., 2015) was built using a lexicon-based NLP approach and later updated to include a wide-array of disambiguation methods.

The dataset was split by the main entity types (Person, Organization, Location), and German DBpedia links were used for the evaluations. We have only taken into account the entities that were linked in the gold standard, therefore

[3] https://dbpedia-spotlight.github.io/demo/
[4] http://babelfy.org/
[5] http://dexter.isti.cnr.it/

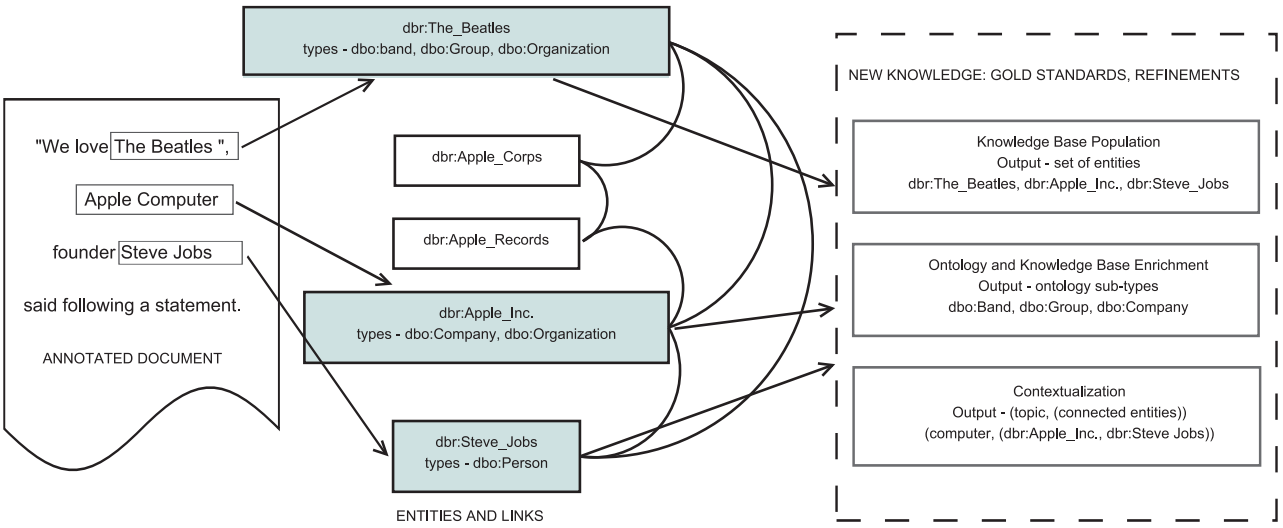

Figure 2: Methods for generating gold standards for knowledge evaluation tasks from existing corpora

no NIL entities were present in this evaluation. Where the German DBpedia links were not available, the owl:sameAs links were taken from the English DBpedia entities, as such links were available for most of the entities from the dataset. Recognyze was only included in the Person and Organisation evaluation, since its Location profiles link entities to GeoNames rather than DBpedia. Many of the Geonames entities are not necessarily linked to German DBpedia and in many cases they are also not linked to the corresponding English DBpedia entity, therefore a direct comparison would have not been possible. The evaluations were performed using the public endpoints of these tools. Using the NIF version of the dataset, GERBIL evaluations can also be performed relatively easy, but only for all the entities and not split by type as this functionality is currently not provided by the framework.

As can be seen from the results, no annotator managed to correctly assign more than half of the entities of the three main entity types (Person, Organisation, Location). Compared to evaluations that use the full datasets (with or without NILs), evaluations of single types provide better insight into the component's performance. It is not uncommon to see differences of several percents after running the experiments again several days later with the same annotators. As outlined in Table 3 all the top tools for a particular entity type are relatively close in terms of F1 measure, although the differences between the types (Person, Organization and Location) are quite significant. Tools that draw upon advanced disambiguation techniques (Babelfy and Recognyze) tend to show higher precision than recall values. These results underline that NEL is a very dynamic field, where most of the evaluated tools outperform their competitors in at least one of the evaluations. We can also conclude that the German language still poses some challenges for the current generation of annotators, especially when comparing these results with those obtained on English corpora (Usbeck et al., 2015).

5 Conclusions

Corpus and dataset updates. With the rise of infrastructure for performing large-scale NLP evaluations (NIF, GERBIL) with multiple tools in parallel comes the need to keep the datasets up to date. We started a discussion with the NIF and GERBIL developers on the steps necessary in order to ensure that datasets are updated to the latest available version of the knowledge base (e.g., tag each entity also with the German DBpedia version, update frequency, etc.). In general we have agreed on a minimal set of maintenance tasks that anyone who publishes a dataset should do (e.g., keep links updated, specify DBpedia version) and some thoughts about this process were published on the forum. It is important to keep datasets updated especially because new entities are always added to DBpedia and therefore some of the entities previously marked as NIL can already have corresponding links in the newer versions. Keeping datasets updated is enabled by the fact that today datasets are shared via GitHub (Braşoveanu et al., 2016).

The evaluation results will help to continuously refine Recognyze, especially the disambiguation algorithms and the query profiles. Having access to a large number of test cases via the documents already annotated while continuously reducing the number of false positives are key to our strategy for improving the NEL performance of Recognyze.

Future versions of this corpus, as well as of other corpora that we may publish, will include more major types in addition to the classic ones, as well as subtypes. In addition to sharing the current corpora, we plan to update the links whenever new versions of DBpedia are available, and repeat the evaluations with these versions. Only some of the tools we have tested were focused on graph disambiguation, therefore, the gold standards for graph disambiguation and relation extraction will be evaluated in future work. In addition, we will address the issue of automated large-scale evaluation in a graph context. Since all the annotator services are available as REST APIs, and some of them are also already integrated in GERBIL,

extracting additional gold standards from existing corpora would open the door to automated large-scale evaluation of graph disambiguation algorithms.

6 Acknowledgements

This work was partially supported by the European Union's 7th Framework Programme via the projects LinkedTV (GA 287911) and DecarboNet (GA 610829), as well as by the Swiss Commission for Technology and Innovation (CTI) via the IMAGINE project (www.htwchur.ch/imagine). The authors wish to thank Giuseppe Rizzo, Raphaël Troncy, Michael Röder and Max Göbel for their valuable feedback.

7 Bibliographical References

Berners-Lee, T., Bizer, C., and Heath, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22.

Bontcheva, K. and Rout, D. (2012). Making Sense of Social Media Streams Through Semantics: A Survey. *Semantic Web*, 1:1–31.

Bunescu, R. C. and Pasca, M. (2006). Using Encyclopedic Knowledge for Named Entity Disambiguation. In *EACL*, volume 6, pages 9–16.

Cattoni, R., Corcoglioniti, F., Girardi, C., Magnini, B., Serafini, L., and Zanoli, R. (2012). The KnowledgeStore: an Entity-Based Storage System. pages 3639–3646. European Language Resources Association (ELRA).

Ceccarelli, D., Lucchese, C., Orlando, S., Perego, R., and Trani, S. (2013). Dexter: An Open Source Framework for Entity Linking. In Paul N. Bennett, et al., editors, *ESAIR'13, Proceedings of the Sixth International Workshop on Exploiting Semantic Annotations in Information Retrieval, co-located with CIKM 2013, San Francisco, CA, USA, October 28, 2013*, pages 17–20. ACM.

Cunningham, H., Maynard, D., and Bontcheva, K. (2011). *Text Processing with GATE*. Gateway Press CA.

Daiber, J., Jakob, M., Hokamp, C., and Mendes, P. N. (2013). Improving Efficiency and Accuracy in Multilingual Entity extraction. In Marta Sabou, et al., editors, *I-SEMANTICS 2013 - 9th International Conference on Semantic Systems, ISEM '13, Graz, Austria, September 4-6, 2013*, pages 121–124. ACM.

Hellmann, S., Stadler, C., and Lehmann, J. (2012). The German DBpedia: A Sense Repository for Linking Entities. In Christian Chiarcos, et al., editors, *Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata*, pages 181–190. Springer.

Hellmann, S., Lehmann, J., Auer, S., and Brümmer, M. (2013). Integrating NLP using Linked Data. In Harith Alani, et al., editors, *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II*, volume 8219 of *Lecture Notes in Computer Science*, pages 98–113. Springer.

Hoffart, J., Yosef, M. A., Bordino, I., Fürstenau, H., Pinkal, M., Spaniol, M., Taneva, B., Thater, S., and Weikum, G. (2011). Robust Disambiguation of Named Entities in Text. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 782–792. Association for Computational Linguistics, ACL.

Ji, H., Dang, H., Nothman, J., and Hachey, B. (2014). Overview of TAC-KBP 2014 Entity Discovery and Linking Tasks. In *Proc. Text Analysis Conference (TAC2014)*.

Navigli, R. and Moro, A. (2014). Multilingual Word Sense Disambiguation and Entity Linking. In *COLING 2014, 25th International Conference on Computational Linguistics, Tutorial Abstracts, August 23-29, 2014, Dublin, Ireland*, pages 5–7. ACL.

Nuzzolese, A. G., Gentile, A. L., Presutti, V., Gangemi, A., Garigliotti, D., and Navigli, R. (2015). Open Knowledge Extraction Challenge. In Fabien Gandon, et al., editors, *Semantic Web Evaluation Challenges. Second SemWebEval Challenge at ESWC 2015, Portoroz, Slovenia, May 31 - June 4, 2015, Revised Selected Papers*, volume 548 of *Communications in Computer and Information Science*, pages 3–15. Springer International Publishing.

Scharl, A., Herring, D., Rafelsberger, W., Hubmann-Haidvogel, A., Kamolov, R., Fischl, D., Föls, M., and Weichselbraun, A. (2016a). Semantic systems and visual tools to support environmental communication. *IEEE Systems Journal*, page (Forthcoming).

Scharl, A., Weichselbraun, A., Göbel, M., Rafelsberger, W., and Kamolov, R. (2016b). Scalable Knowledge Extraction and Visualization for Web Intelligence. In *49th Hawaii International Conference on System Sciences (HICSS-2016)*, pages 3749–3757. IEEE.

Usbeck, R., Röder, M., Ngomo, A. N., Baron, C., Both, A., Brümmer, M., Ceccarelli, D., Cornolti, M., Cherix, D., Eickmann, B., Ferragina, P., Lemke, C., Moro, A., Navigli, R., Piccinno, F., Rizzo, G., Sack, H., Speck, R., Troncy, R., Waitelonis, J., and Wesemann, L. (2015). GERBIL: General Entity Annotator Benchmarking Framework. In Aldo Gangemi, et al., editors, *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pages 1133–1143. ACM.

Weichselbraun, A., Streiff, D., and Scharl, A. (2015). Consolidating Heterogeneous Enterprise Data for Named Entity Linking and Web Intelligence. *International Journal on Artificial Intelligence Tools*, 24(2):1540008, 1–31.

8 Language Resource References

Braşoveanu, A.M.P. and Nixon, L.J.B. and Weichselbraun, A., and Scharl, A. (2016). *RBB150*. LinkedTV GitHub Account (www.github.com/linkedtv/videocorpus), 1.0.