PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 108 JUNE 2017 355–366

Improving Machine Translation through Linked Data

Ankit Srivastava, Georg Rehm, Felix Sasaki

German Research Center for Artificial Intelligence (DFKI),
Language Technology Lab, Berlin, Germany

Abstract

With the ever increasing availability of linked multilingual lexical resources, there is a renewed interest in extending Natural Language Processing (NLP) applications so that they can make use of the vast set of lexical knowledge bases available in the Semantic Web. In the case of Machine Translation, MT systems can potentially benefit from such a resource. Unknown words and ambiguous translations are among the most common sources of error. In this paper, we attempt to minimise these types of errors by interfacing Statistical Machine Translation (SMT) models with Linked Open Data (LOD) resources such as DBpedia and BabelNet. We perform several experiments based on the SMT system Moses and evaluate multiple strategies for exploiting knowledge from multilingual linked data in automatically translating named entities. We conclude with an analysis of best practices for multilingual linked data sets in order to optimise their benefit to multilingual and cross-lingual applications.

1. Introduction

Statistical Natural Language Processing (NLP) technologies rely on large volumes of data from which models can be constructed to leverage patterns and knowledge from these data sets. Typically, these resources are in the form of annotated (structured, labeled) or unstructured natural language text such as aligned input and output language paired sentences for Machine Translation (MT) or parsed treebanks for parsing. However, we can observe a certain shortage of NLP systems (Nebhi et al., 2013; Hokamp, 2014) which exploit knowledge from structured or semi-structured resources such as the Linked Open Data (LOD) lexical resources created for and maintained as part of the Semantic Web and its Linked Data Cloud. This shortage is most likely due to the fact that the MT community is primarily focused upon con-

 Corresponding author: ankit.srivastava@dfki.de
Cite as: Ankit Srivastava, Georg Rehm, Felix Sasaki. Improving Machine Translation through Linked Data. The Prague Bulletin of Mathematical Linguistics No. 108, 2017, pp. 355–366. doi: 10.1515/pralin-2017-0033.

tinuously improving their respective rule-based, statistical or neural algorithms and approaches, while the LOD community is focused upon representing, providing and linking data sets. Our contribution is an approach at building a bridge between the two communities.

In this paper, using Statistical Machine Translation (SMT) as a case-study, we explore three strategies for leveraging knowledge from a variety of LOD resources. In addition to analysing the impact of linked data on MT, we briefly discuss considerations for creating and linking multilingual lexical resources on the web so that NLP systems can benefit from them.

This paper is structured as follows. We briefly overview the background technologies (Semantic Web, Resource Description Format, Linked Open Data, SMT workflow) leveraged in this research in Section 2. In Section 3, we outline three strategies for integrating linked data in a SMT system followed by a summary of previous works in Section 4. In Sections 5 and 6 we describe our experimental results and analysis after which we conclude in Section 7.

2. Technologies

In this section, we briefly summarise the technologies used, i. e., Statistical Machine Translation (SMT), Linked Open Data (LOD) resources, and Semantic Web technologies facilitating the integration of SMT with LOD.

2.1. Semantic Web Technologies

With regard to the Semantic Web, several key technologies can be exploited in NLP systems.[1]

RDF (Resource Description Framework) is a formalism to represent data on the web as a labelled graph of triples (subject, predicate, object, or, to put it another way, objects and their relations). URIs (Uniform Resource Identifiers) are compact sequences of characters used to identify resources – including objects – on the web. Ontologies are hierarchical vocabularies of types and relations, allowing more efficient storage and use of data by encoding generic facts about objects. RDF Schema (RDFS) is one such formalism or knowledge representation language, OWL (Web Ontology Language) can be used to represent more complex knowledge structures. RDF and RDFS are the underlying syntax and ontology as well as vocabulary languages, used to represent machine readable data and define relevant properties such as rdfs:label for language name. SPARQL[2] is the query language used to retrieve information from RDF-encoded data including NIF.

[1]The basic technologies, data formats and approaches that constitute the technological building blocks of the Semantic Web and Linked Data are developed and standardised by the World Wide Web Consortium (W3C).

[2]http://www.w3.org/TR/rdf-sparql-query/

The knowledge sources employed in our experiments are structured as Linked Data, stored in RDF (subject-predicate-object triples). In order to access or retrieve information (translations) from the RDF datasets for integration in a MT system, we need to query the database using SPARQL. The example below illustrates a sample SPARQL query for retrieving the German (de) translation of the term "Microsoft Paint."

Listing 1. An example SPARQL query

```
PREFIX dbpedia: <http://dbpedia.org/resource/>

SELECT distinct *
WHERE {
    <http://dbpedia.org/resource/Paint_(software)>
    rdfs:label ?label
    filter langMatches( lang(?label), "de" )
}
```

NIF 2.0[3] (Natural Language Processing Interchange Format) is an RDF-based format that aims to achieve interoperability between NLP tools such as parsers, SMT engines and annotated language resources such as DBpedia. Its joint application with technologies like ITS 2.0[4] (Internationalization Tag Set) and the OntoLex lemon model[5] makes it an ideal candidate to implement multilingual applications. The primary use case of this standard is to serve as an input and output format for web services, that enable seamless pipelining or combination of various language and linked data processing web services in sequence. With regard to NLP, an important characteristic of NIF is that its atomic unit is a character rather than a word. Thus, if a sentence has 23 characters (including spaces between words), the resource or sentence spans from 0 to 22. In this way, NLP pipelines can create fine grained annotations relying on the graph based model of RDF. In order to evaluate effectiveness of LOD in SMT (primary aim of this paper), we integrated our SMT system with NIF input / output wrappers using methodology described in (Srivastava et al., 2016).

In a nutshell, if the data such as a multilingual lexicon is stored as a linked data (NIF / RDF), then SPARQL is a tool to retrieve information from the linked data such as translations in the required target language.

[3]http://persistence.uni-leipzig.org/nlp2rdf/

[4]http://www.w3.org/TR/its20/

[5]https://www.w3.org/2016/05/ontolex/

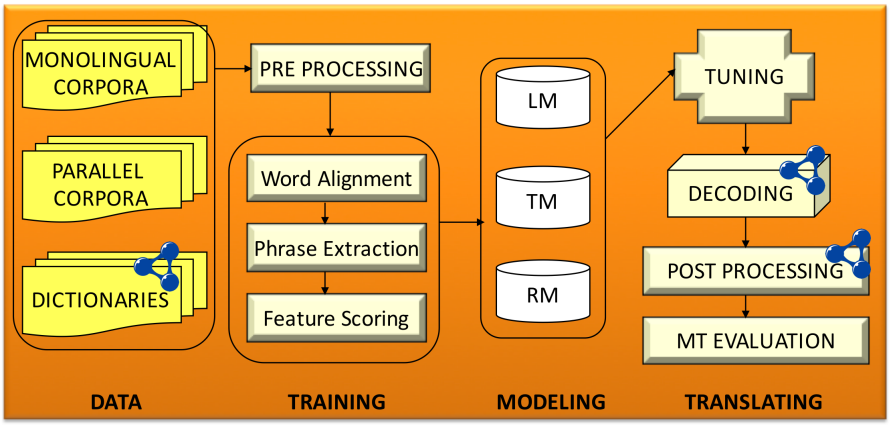

Figure 1. Workflow of the SMT modules

2.2. Machine Translation

There has been an ever increasing interest in Machine Translation, one of the earliest non-numeric applications of computers (Hutchins, 2000), since the enormous increase of multilingual user-generated content on the web. There are a number of approaches to implementing MT (rule-based, example-based, statistical, syntax-based, semantics-based, hybrid, and neural). Statistical MT is the most widely researched paradigm and represents, along with neural MT, the current state of the art.

In this paper, we conduct our experiments using the widely deployed open source SMT system Moses[6] (Koehn et al., 2007). We use the phrase-based SMT system with standard configurations, as specified in Section 5. Several similar approaches exist such as an attempt on integrating bilingual dictionaries in SMT (Arcan et al., 2014).

Figure 1 shows the workflow of a typical SMT system. Data in the form of bilingual (including dictionaries extracted from LOD) and monolingual resources (typically collections of documents tokenised into sentences) is fed into the training network which creates the language model (LM), translation model (TM), and reordering model (RM). These models are then tuned, followed by decoding (the actual translation step), followed by post-processing (such as linked data translation edits).

As shall be described in Section 3, we integrate linked data (illustrated with the blue triangular structure in Figure 1) into the SMT system at three points:

- As dictionaries during training (before Word Alignment)
- As an alternate resource (translation rules) during decoding

[6]http://www.statmt.org/moses/

- As rules in the form of post-editing process

2.3. Linked Data Resources

For the MT improvement, we are going to use three linked data resources: DBpedia, BabelNet, and JRC-names. These three resources are part of the Linguistic Linked Open Data Cloud[7], an interconnected set of linguistic resources represented as linked data. The LLOD cloud helps to address problems in various research and application areas, such as interoperability of linguistic annotations, graph-based annotations based on the linked data graph model without the need for special purpose tools, or fast increase of multilingual resources via ease of linkage. In addition, based on LLOD principles, new formats like OntoLex (Fiorelli et al., 2015) have been put forward.

DBpedia[8] is a linked open dataset (extracted from Wikipedia) consisting of 4.58 million entities in up to 125 languages and 29.8 million links to external web pages. DBpedia has been used in many linked data applications. For the improvement of MT it is useful because of the high number of multilingual labels, and the high number of cross-lingual links between DBpedia instances. DBpedia Spotlight[9] is an open-source tool for automatically annotating mentions of DBpedia resources in text. Note that the translations may be prone to error on account of being user generated.

BabelNet (Navigli and Ponzetto, 2012) is a a multilingual resource created by linking Wikipedia to WordNet and other semantic networks, filling gaps with MT. BabelNet is highly multilingual and, since it encompasses, e. g., DBpedia, we expect an additional improvement of MT compared to using DBpedia only.

JRC-names[10] (Steinberger et al., 2011) is a freely available multilingual named entity resource for person and organisation names that have been compiled from over seven years of analysing multilingual news articles. Since March 2016, JRC-Names has also been available as linked data, including additional information such as frequencies per language, titles found with the entities, and date ranges.

Table 1 gives a comparative evaluation of the languages and sizes of these three resources.

3. Integrating LOD into SMT – Three Approaches

As regards integrating Linked Open Data resource into Machine Translation workflows, we implemented three different strategies (illustrated in Figure 1).

[7]http://linguistic-lod.org/llod-cloud

[8]http://wiki.dbpedia.org

[9]https://github.com/dbpedia-spotlight/

[10]https://ec.europa.eu/jrc/en/language-technologies/jrc-names

Resource	# Entries	# Languages
DBpedia	23.8 million	125
BabelNet	14 million	270
JRC-Names	205 thousand	22

Table 1. Comparison of Linked Data resources

- **Dictionaries:** Transform LD resources into a dictionary for word alignment such that the models will contain knowledge from the Linked Data resource and let the Moses decoder decide which translation knowledge (linked data or parallel corpora) to retrieve.
- **Pre-decoding:** Forced decoding by first named entity linking via SPARQL query (using Moses xml-input exclusive feature).
- **Post-processing:** Automatic post-editing or correcting of untranslated words, i.e. translations which are not present in the translation model.

Note that each of the three algorithms are applied individually to each of the three LOD resources (DBpedia, BabelNet, JRC-named), described in Section 5.

3.1. Algorithm 1: Dictionaries

Each of our LOD resources (DBpedia, BabelNet, JRC-names) is available as a bilingual dictionary on their respective websites. For the dictionary approach, we treat these dictionaries as an additional bilingual terminology dataset and integrate them into the SMT system using well-known methods of adding bilingual terms to the training data before the word and phrase alignment step of training (Bouamor et al., 2012).

3.2. Algorithm 2: Pre-decoding

The term pre-decoding alludes to the fact that the LOD resource is gathered right before calling the SMT decoder. In reality, the linked data resource provides additional translation rules for specific words and phrases (mainly named entities) during decoding. The pre-decoding algorithm inspired by the approach in (Srivastava et al., 2016) is described below:

1. Take as input a source sentence
2. Tag the named entities using an off-the-shelf Named Entity Recogniser
3. For each of the named entities invoke a SPARQL query for the appropriate resource (DBpedia, JRC-names, BabelNet) to retrieve the translation in the target language

4. Integrate these translations in the Moses decoder. Encode the named entity and its translation in a format compatible with the Moses decoder (enabled with the xml-input feature)

Note that all the procedures above are carried out by freely available web service API calls, the source code for which can be found at `https://github.com/freme-project` for FREME web services[11] and at `https://github.com/dkt-projekt` for DKT web services.[12]

3.3. Algorithm 3: Post-processing

As mentioned previously, a major source of error in MT is the presence of unknown words, i.e. entries which do not have a valid translation in the training data. This is particularly true when the SMT system is trained in a domain different from the domain of the test data, as is typical of large-scale evaluations such as the WMT Shared Tasks (Bojar et al., 2016). Our third algorithm identifies the untranslated words[13] and calls a SPARQL query to retrieve the translation (if available) from each of the three LOD resources. The SPARQL Query endpoints are available at:

- DBpedia: `http://de.dbpedia.org/sparql`
- BabelNet: `http://babelnet.org/sparql/`
- JRC-names: `http://data.europa.eu/euodp/en/linked-data`

4. Related Work

We use multiple linked data resources using three different strategies. There have been previous attempts at integrating LOD into SMT, however, to the best of our knowledge, none of these demonstrated all approaches on one dataset like we do in this subsmission. (McCrae and Cimiano, 2013) primarily integrated the dictionary of translations extracted from LOD resources during decoding and created a new feature for linked data. They essentially let the Moses decoder decide when to choose translations from LOD and when to translate from its phrase tables. In contrast to our approach on encoding documents in NIF (while entity linking via SPARQL queries), they employ another ontology called Lemon (Lexicon Model for Ontologies[14]) to translate unknown words, i. e., translations not found by the decoder. Our Algorithm 1 (Dictionaries) is most similar to their approach while we employ an alternative approach to handling unknown words (Algorithm 3 [Post-processing]).

(Du et al., 2016) extracted translations from BabelNet dictionaries using both (McCrae and Cimiano, 2013)'s methods as well as the post processing (Algorithm 3)

[11] Of particular interest is the web service named e-entity/dbpedia-spotlight.

[12] Of particular interest are the services DKTBrokerStandalone/nifTools, e-NLP/Sparqler, and e-SMT.

[13] Moses allows special annotation to highlight the presence of unknown words in the translated output

[14] http://lemon-model.net

Category	Training	Development	Test
Dataset	Europarl v7	newstest 2011	newstest 2012
German–English	1,920,209	3,003	3,003
Spanish–English	1,965,374	3,003	3,003

Table 2. Statistics of parallel corpus used in baseline SMT training experiments

System	BLEU	TER
Baseline	12.30	0.788
DBpedia	12.33	0.776
BabelNet	12.25	0.780
JRC-Names	12.39	0.762

Table 3. Evaluation results on English–German

employed in this contribution to demonstrate modest improvements in translating English-Polish and English-Chinese data.

The pre-decoding approach of locating named entities and forcing their translations from LOD resources (retrieved via SPARQL queries) on to the decoder was inspired by methodology described in (Srivastava et al., 2016).

5. Experiments

We trained the SMT system to translate from English to German and Spanish. The set of parallel sentences for training, and the development and test sets for tuning and testing respectively were sourced from the data provided for the WMT 2012 shared task on MT[15]. This was done mainly to make our experiments comparable to that of (McCrae and Cimiano, 2013). Table 2 gives an overview of the data sizes our models are trained on.

Tables 3 and 4 show the evaluation results of our MT experiments. The Baseline system did not use any linked data of any sort. The two evaluation metrics used are BLEU (Papineni et al., 2002) and TER (Snover et al., 2006).

Contrary to our expectation, BabelNet did not perform as well as other linked data resources. While JRC-Names gave the best performance, probably owing to their data being from the same domain as the test data (news domain). We also believe that

[15]http://www.statmt.org/wmt12/

System	BLEU	TER
Baseline	31.70	0.577
DBpedia	31.03	0.550
BabelNet	30.99	0.558
JRC-Names	31.91	0.540

Table 4. Evaluation results on English–Spanish

BabelNet being the largest resource in terms of size also contained more noise and it was often difficult to disambiguate translations.

6. Analysis of Multilingual Linked Data Sets

Compared to previous approaches, see (McCrae and Cimiano, 2013), our experiments do not provide a high improvement of MT quality. However, we can draw useful conclusions in light of best practices for creating linguistic LOD. The forum for the best practices is the BPMLOD Community Group, see[16]. We examined guidelines in the realm of BPMLOD, for linguistic linked data resources such as BabelNet[17] and bilingual dictionaries[18]. Based on the experiments we conducted, there are a few features which are of importance for applying linguistic LOD in MT.

- Domain Identifier: When a specific term has multiple translations in another language, properties such as the domain would help in disambiguating the context.
- Morphology: When translating into a morphologically richer language, information about the form of a noun changes based on the case can help to improve the translation quality.

In conducting a manual evaluation of the results i.e. having a bilingual German speaker eye a randomly selected subset of the translated outputs, we also discovered that while our systems are useful in disambiguating erroneous translations, the automatic MT evaluation metrics are deficient in counting them such that they do not account for variability in translations. For example the reference translation "MS Paint" only matches partially with the LOD system translation "Microsoft Paint." Algorithm 2 (Pre-decoding) identified and correctly translated 15% more terms (named entities) than the baseline SMT system.

[16]https://www.w3.org/community/bpmlod/

[17]http://www.w3.org/2015/09/bpmlod-reports/multilingual-dictionaries/

[18]http://www.w3.org/2015/09/bpmlod-reports/bilingual-dictionaries/

Source (en):	**MS Paint** is a good option.
Baseline (de):	**Frau Farbe** ist eine gute wahl.
Linked (de):	**Microsoft Paint** ist eine gute wahl.
Reference (de):	**MS Paint** ist eine gute Möglichkeit.

7. Conclusion and future work

In this paper, we demonstrated employing knowledge from three semantic web resources which show modest improvement in English-German and English-Spanish translations. We leave for future work exploiting several more features such as word senses from the knowledge-rich semantic network in MT.

While deep learning-based neural approaches to MT (i. e., NMT: Neural Machine Translation (Sennrich et al., 2016)) have been the state of the art since WMT 2016, we decided to demonstrate our Linked Data-focused approach using SMT due to the lower complexity of the integration task. Future work will include experiments with NMT using our Linked Data-focused approach at improving MT systems. Note that the post-process (Algorithm 3) approach can be theoretically applied to a neural MT system as-is.

It is our belief that the use of Linked Open Data in combination with Named Entity Recognition (Algorithm 2 [Pre-decoding] in our approach) helps reduce the long tail of difficult to translate names. This is similar to word sense disambiguation in MT (Carpuat, 2008). Employing world knowledge for disambiguating terms other than named entities is another potential direction for future research.

This paper is a step towards making MT semantic web-aware and it is our hope that more MT researchers undertake integration of this fertile knowledge source.

Acknowledgements

We would like to thank the anonymous reviewers for their insightful and helpful comments. The project Digitale Kuratierungstechnologien (DKT) is supported by the German Federal Ministry of Education and Research (BMBF), Unternehmen Region, instrument Wachstumskern-Potenzial (No. 03WKP45). More information on the project can be found online at `http://www.digitale-kuratierung.de`.

Bibliography

Arcan, Mihael, Marco Turchi, Sara Tonelli, and Paul Buitelaar. Enhancing Statistical Machine Translation with bilingual terminology in a CAT environment. In *11th Conference of the Association for Machine Translation in the Americas*, pages 54–68, 2014.

Bojar, Ondrej, Christian Buck, Rajen Chatterjee, Christian Federmann, Liane Guillou, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Auralie Navaol, Mariana Neves, Pavel Pecina, Martin Popel, Philipp Koehn, Christof Monz, Matteo Negri, Matt Post, Lucia Specia, Karin Verspoor, Jorg Tiedemann, and Marco Turchi, editors. *Proceedings of the First Confer-*

ence on Machine Translation. Association for Computational Linguistics, Berlin, Germany, August 2016. URL http://www.aclweb.org/anthology/W/W16/W16-2200.

Bouamor, Dhouha, Nasredine Semmar, and Pierre Zweigenbaum. Identifying bilingual Multi-Word Expressions for Statistical Machine Translation. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012)*, pages 674–679, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7. URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/886_Paper.pdf. ACL Anthology Identifier: L12-1527.

Carpuat, Marine Jacinthe. *Word Sense Disambiguation for Statistical Machine Translation*. PhD thesis, 2008. AAI3350676.

Du, Jinhua, Andy Way, and Andrzej Zydron. Using BabelNet to Improve OOV Coverage in SMT. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 2016.

Fiorelli, Manuel, Armando Stellato, John P. Mccrae, Philipp Cimiano, and Maria Teresa Pazienza. LIME: The Metadata Module for OntoLex. In *Proceedings of the 12th European Semantic Web Conference on The Semantic Web. Latest Advances and New Domains - Volume 9088*, pages 321–336, New York, NY, USA, 2015. Springer-Verlag New York, Inc. ISBN 978-3-319-18817-1. doi: 10.1007/978-3-319-18818-8_20. URL http://dx.doi.org/10.1007/978-3-319-18818-8_20.

Hokamp, Chris. Leveraging NLP Technologies and Linked Open Data to Create Better CAT Tools. In *International Journal of Localisation, Vol 14*, pages 14–18, 2014.

Hutchins, John. *John W. Hutchins (Eds.), Early Years in Machine Translation*, chapter The first decades of Machine Translation: overview, chronology, sources, pages 1–16. John Benjamins B. V., 2000.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1557769.1557821.

McCrae, John and Philipp Cimiano. Mining Translations from the Web of Open Linked Data. In *Proceedings of the Joint Workshop on NLP, LOD and SWAIE*, pages 8–11, 2013.

Navigli, Roberto and Simone Paolo Ponzetto. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. In *Artificial Intelligence*, pages 217–250, 2012.

Nebhi, Kamel, Luka Nerima, and Eric Wehrli. NERTIS - A Machine Translation Mashup System using Wikimeta and DBpedia. In *Semantic Web (ESWC) 2013 Satellite Events*, pages 312–318, 2013.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jung Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. Edinburgh Neural Machine Translation Systems for WMT 16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany, August 2016. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/W/W16/W16-2323`.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciula, and John Makhoul. A Study of Translation Edit Rate with targeted Human Annotation. In *7th Conference of the Association for Machine Translation in the Americas*, pages 223–231, 2006.

Srivastava, Ankit, F. Sasaki, P. Bourgonje, J. Moreno-Schneider, J. Nehring, and G. Rehm. How To Configure Statistical Machine Translation with Linked Open Data Resources. In *Proceedings of the 38th Annual Translating and Computer Conference*, TC 38, 2016.

Steinberger, Ralf, Bruno Pouliquen, Mijail Kabadjov, Jenya Belyaeva, and Erik van der Goot. JRC-NAMES: A Freely Available, Highly Multilingual Named Entity Resource. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 104–110. Association for Computational Linguistics, 2011. URL `http://aclweb.org/anthology/R11-1015`.

Address for correspondence:
Ankit Srivastava
`ankit.srivastava@dfki.de`
DFKI GmbH
Alt-Moabit 91c
10559 Berlin, Germany