DOCTORAL DISSERTATION

Knowledge Graphs for Multilingual Language Translation and Generation

A dissertation presented
by
Diego Campos Moussallem
to the
Faculty for Computer Science,
Electrical Engineering and Mathematics
of
Paderborn University

in partial fulfillment of the requirements
for the degree of
Dr. rer. nat.

Paderborn, Germany
March 2020

DISSERTATION

Knowledge Graphs for Multilingual Language Translation and Generation
Diego Campos Moussallem, Paderborn University
Paderborn, Germany, 2020

REVIEWERS

Prof. Dr. Axel-Cyrille Ngonga Ngomo, Paderborn University
Prof. Dr. Sören Auer, Leibniz Universität Hannover
Prof. Dr. Jens Lehmann , Universität Bonn

DOCTORAL COMMITTEE

Prof. Dr. Axel-Cyrille Ngonga Ngomo, Paderborn University
Prof. Dr. Sören Auer, Leibniz Universität Hannover
Prof. Dr. Jens Lehmann, Universität Bonn
Prof. Dr. Heike Wehrheim, Paderborn University
Prof. Dr. Gregor Engels, Paderborn University

Abstract

KNOWLEDGE GRAPHS FOR MULTILINGUAL LANGUAGE TRANSLATION AND GENERATION

The Natural Language Processing (NLP) community has recently seen outstanding progress, catalysed by the release of different Neural Network (NN) architectures. Neural-based approaches have proven effective by significantly increasing the output quality of a large number of automated solutions for NLP tasks (Belinkov and Glass, 2019). Despite these notable advancements, dealing with entities still poses a difficult challenge as they are rarely seen in training data. Entities can be classified into two groups, i.e., proper nouns and common nouns. Proper nouns are also known as Named Entities (NE) and correspond to the name of people, organizations or locations, e.g., *John*, *WHO* or *Canada*. Common nouns describe classes of objects, e.g., *spoon* or *cancer*. Both types of entities can be found in a Knowledge Graph (KG). Recent work has successfully exploited the contribution of KGs in NLP tasks, such as Natural Language Inference (NLI) (K M et al., 2018) and Question Answering (QA) (Sorokin and Gurevych, 2018). Only a few works had exploited the benefits of KGs in Neural Machine Translation (NMT) when the work presented herein began. Additionally, few works had studied the contribution of KGs to Natural Language Generation (NLG) tasks. Moreover, the multilinguality also remained an open research area in these respective tasks (Young et al., 2018).

In this thesis, we focus on the use of KGs for machine translation and the generation of texts to deal with the problems caused by entities and consequently enhance the quality of automatically generated texts. Before handling entities in translation or generation, the first research challenge of this thesis lies in the disambiguation of entities. Some entities are highly ambiguous, e.g., *Kiwi* can be a fruit or bird. However, once they are disambiguated, their translations are found in multilingual KGs. We addressed this challenge by devising MAG, a multilingual knowledge graph-based entity linking approach for 40 languages. MAG achieves an average of 0.63 F-measure across all languages and places first out of 13 annotation systems.

Our second research challenge is how to cope with entities while generating natural language sentences in different languages from Resource Description Framework (RDF) KGs. The underlying rationale is that generating entities from KGs shares similar NMT problems as translating them between languages in texts. We noticed that previous work has predominantly focused on English, and only a few works provided solutions for other languages. We dealt with this challenge by creating a Portuguese RDF verbalizer, named RDF2PT, which was further extended to Spanish and English. RDF2PT generates sentences and small summaries in Portuguese, which show fluency almost equivalent to humans, scoring 4 (exact mean) on a 5-Likert scale. Further, we examined the Referring Expression Generation (REG) task that aims to choose the referential form of entities while generating texts. We then created the first neural-based REG model, named NeuralREG, which clearly outperforms the state of the art , scoring 5.26 (exact mean) on a 7-Likert scale.

Our third research challenge involves the translation of entities in text. With this aim, we applied KGs into NMT models. We thus created the first KG-augmented NMT model, named KG-NMT, by combining Entity Linking (EL) and Knowledge Graph Embeddings (KGE). KG-NMT achieves consistent translation improvements up to +3 BLEU, METEOR, and chrF3 on open domain datasets, and on domain-specific data and ontologies. Later, we discerned that applying KGs into NLP tasks requires rich language-based KGs. We therefore devised our fourth research challenge which pertains to the low resource language problem in KGs. To that end, we developed the first neural-based approach, named THOTH, for translating and enriching KGs across languages. THOTH achieves a translation accuracy of 86%, and its artificially enriched KGs improve the EL task by +19% F-measure. Overall, our findings show that the application of KGs is an effective way of handling entities and addressing its related data sparsity issues in multilingual text translation and generation.

Zusammenfassung

KNOWLEDGE GRAPHS FOR MULTILINGUAL LANGUAGE TRANSLATION AND GENERATION

Die Natural Language Processing (NLP)-Gemeinschaft hat in letzter Zeit herausragende Fortschritte erzielt, die durch die Veröffentlichung verschiedener Architekturen künstlicher neuronaler Netze (NN) katalysiert wurden. NN-basierte Ansätze haben sich als effektiv erwiesen, da sie die Qualität der automatisiert erstellten Lösungen für eine große Zahl von NLP-Aufgaben (Belinkov and Glass, 2019) deutlich erhöht haben. Trotz dieser bemerkenswerten Fortschritte stellt der Umgang mit Entitäten immer noch eine schwierige Herausforderung dar, da sie in den Trainingsdaten nur selten zu vorkommen. Entitäten lassen sich in zwei Gruppen einteilen: Eigennamen und Gattungsnamen. Eigennamen werden auch als Named Entities (NE) bezeichnet und entsprechen den Namen von Personen, Organisationen oder Orten, z. B. *John*, *WHO* oder *Kanada*. Gattungsnamen beschreiben Klassen von Objekten, z. B. *Löffel* oder *Krebs*. Beide Typen von Entitäten können in einem Wissensgraphen (KG) gefunden werden. In jüngster Zeit wurden KGs erfolgreich bei der Lösung von NLP-Aufgaben genutzt, wie z. B. Natural Language Inference (K M et al., 2018) und Question Answering (Sorokin and Gurevych, 2018). Dagegen haben sich nur wenige Arbeiten mit der Anwendung von KGs für die maschinelle, neuronale Übersetzung (NMT) oder der Generierung von natürlicher Sprache (NLG) beschäftigt, als mit dieser Arbeit begonnen wurde. Darüber hinaus ist die Mehrsprachigkeit bei diesen beiden Problemen weiterhin ein offenes Forschungsgebiet (Young et al., 2018).

In dieser Arbeit konzentrieren wir uns auf die Verwendung von KGs für die maschinelle Übersetzung und die Generierung von Texten, um die durch Entitäten verursachten Probleme zu behandeln und folglich die Qualität automatisch generierter Texte zu verbessern. Zuvor wird in dieser Arbeit die Disambiguierung von Entitäten behandelt. Einige Entitäten sind hochgradig mehrdeutig, z. B. kann es sich bei dem Begriff *Kiwi* um eine Frucht oder einen Vogel handeln. Die Disambiguierung ermöglicht letztendlich das Auffinden von Übersetzungen in mehrsprachigen KGs. Zur Auflösung von Ambiguitäten

wurde das Framework MAG entwickelt, welches auf mehrsprachigen Wissensgraphen basiert und die Verknüpfung von Entitäten in über 40 Sprachen ermöglicht. MAG erreicht ein durchschnittliches F-Measure von 0,63 über alle 40 Sprachen und steht damit an erster Stelle von 13 Annotationssystemen.

Im zweiten Teil wird die Frage behandelt, wie mit Entitäten bei der Generierung von Sätzen basierend auf dem Resource Description Framework (RDF) in verschiedenen natürlichen Sprachen umzugehen ist. Die zugrundeliegende Überlegung ist, dass die Erzeugung von Entitäten aus KGs ähnliche NMT-Probleme aufweist wie die Übersetzung zwischen Sprachen in Texten. Da sich vorherige Ansätze hauptsächlich auf die englische Sprache fokussieren und es nur wenige Ansätze für weitere Sprachen gibt, wurde ein RDF-Verbalizer entwickelt, welcher sowohl portugiesische, spanische als auch englische Texte generieren kann. RDF2PT erzeugt Sätze und kleine Zusammenfassungen auf Portugiesisch, die eine fast menschenähnliche Sprachkompetenz zeigen und auf einer 5-Likert-Skala im Mittel mit 4 bewertet wird. Ferner wurde das Referring Expression Generation (REG) Problem behandelt, welches sich mit der Auswahl der referentielle Form von Entitäten beschäftigt. Anschließend entwickelten wir das erste NN-basierte REG-Modell, genannt NeuralREG, das den Stand der Technik deutlich übertrifft und auf einer 7-Likert-Skala mit 5,26 (exakter Mittelwert) bewertet wurde.

Unsere dritte Forschungsaufgabe betrifft die Übersetzung von Entitäten in Textex. Ziel ist es KGs in NMT-Modelle zu integrieren, indem Entity Linking (EL) und Knowledge Graph Embeddings (KGE) zu einem KG-NMT Modell kombiniert werden. KG-NMT erzielt konsistente Übersetzungsverbesserungen von bis zu +3 BLEU, METEOR und chrF3 bei offenen Domänendatensätzen und bei domänenspezifischen Daten und Ontologien. Zudem wurde festgestellt, dass die Anwendung von KGs in NLP-Aufgaben umfangreiches sprach basiertes KGs erfordert. Folglich beschäftigt sich der vierte Teil dieser Arbeit mit der Behandlung von Sprachen, für die nur wenig sprach basiertes Wissen verfügbar ist. Zu diesem Zweck entwickelten wir den ersten neuronal-basierten Ansatz namens THOTH, um KGs sprachübergreifend zu übersetzen und anzureichern. THOTH erreicht eine Übersetzungsgenauigkeit von 86%, und seine künstlich angereicherten KGs verbessern die EL-Aufgabe um +19% F-Measure. Insgesamt zeigen unsere Ergebnisse, dass die Anwendung von KGs eine effektive Methode ist, um mit Entitäten umzugehen und die damit verbundenen Probleme der Datensparsamkeit bei der Übersetzung und Erstellung mehrsprachiger Texte zu lösen.

Acknowledgments

First of all, I would like to thank my wife, Carol. This thesis has only been possible due to her unconditional support. It is uncountable how often she supported me during all the challenging and stressful moments for completing this thesis. The thesis was written within the Data Science research group (DICE), led by Prof. Dr. Axel-Cyrille Ngonga Ngomo, who I wholeheartedly thank for being my advisor and granting me the freedom to develop and pursue my research ideas that have lead to this thesis work. I want to thank Dr. Sebastian Hellmann for inviting me to become a member of the Agile Knowledge Engineering and Semantic Web (AKSW) group at the University of Leipzig, where DICE began to take shape. I am delighted to be part of AKSW and DICE and have had the opportunity to work with so many talented people. Additionally, I want to thank Dr. Ricardo Usbeck for always supporting me within and out of the Ph.D. environment. I extend my gratitude to Dr. Paul Buitelaar and Dr. Mihael Arcan for the invitation to collaborate and do an internship in their research group, INSIGHT, at the National University of Ireland Galway. I am also grateful to have had the opportunity to meet Dr. Thiago Castro Ferreira, with whom I worked closely in various research projects within the framework of natural language generation. Special thanks go to Dr. Diego Esteves, who introduced me to the ASKW group and encouraged me to pursue a Ph.D. abroad. He became more than a true friend, a brother, and I will always be grateful to him. Moreover, I would like to thank my friends and colleagues from Leipzig and Paderborn, especially Tommaso Soru, André Valdestilhas, Edgard Marx, Kleanthi Georgala, Kunal Jha, and Amrapali Zaveri (in memoriam). It was an honor to meet all of you. Furthermore, I would like to extend my deepest gratitude to my family and to my parents, Vilma and Makhoul, especially my mother, who sacrificed her life to raise me to who I am now. I would also like to thank my mother-in-law, Isabel, for her great support. Next, I would like to thank my English teacher, Marcos Moret, for accompanying me all this time and for becoming a true friend. Finally, I want to thank the National Council for Scientific and Technological Development (CNPq) for the scholarship, which supported a significant part of this research. Likewise, I would like to thank DAAD for funding my German course and travel expenses.

Contents

1 Introduction

The technological progress of recent decades has made both the distribution of and access to content in different languages simpler. Still, the Web has approximately 48% of the pages unavailable in English.[1] Translation aims to support users who need to access content in a language in which they are not fluent (Slocum, 1985; Koehn, 2010).

However, translation is a difficult task due to the complexity and diversity of the natural language families (Jurafsky, 2000). In addition, manual translation does not scale to the magnitude of the Web. One remedy for this problem is Machine Translation (MT). The main goal of MT is to enable people to assess content in languages other than the languages in which they are fluent (Bar-Hillel, 1960). From a formal point of view, this means that the goal of MT is to transfer semantics from a piece of text in an input language to a piece of text in an output language (Hutchins and Somers, 1992). At the time of writing, large information portals such as Google[2] or Bing[3] already offer MT services even though they are not entirely open-source.

MT systems are now popular on the Web, but they still generate a large number of incorrect translations. The two most common types of errors are responsible for roughly 70% of the translation errors: 40% of the translation errors are the result of reordering errors, where an MT system outputs sentences in a target language with incorrect word

[1] https://www.internetworldstats.com/stats7.htm
[2] http://translate.google.com.br/about/
[3] http://www.bing.com/translator/help/

sequence. Another 30% are due to lexical and syntactical ambiguity, i.e., when a single sentence or a word can have more than one meaning (Moussallem et al., 2018c). Thus, addressing these barriers is a key challenge for modern translation systems.

Recently, a novel Statistical Machine Translation (SMT) paradigm has emerged called NMT. NMT relies on NN algorithms. NMT has been achieving significant improvements and is now the state of the art in MT approaches. Since NMT has shown impressive results on reordering (Stahlberg, 2019), an important challenge in NMT lies in the disambiguation process, both at the syntactic and semantic levels. Additionally, NMT approaches struggle with out-of-vocabulary (OOV) words (rare words) since they operate with a fixed vocabulary size. Although the community has been combining efforts to address this problem by proposing character-based (Luong and Manning, 2016; Chung et al., 2016) or Byte Pair Encoding (BPE) models (Sennrich et al., 2016a), OOV words are still an open problem as they are highly co-related to the disambiguation of entities (Koehn and Knowles, 2017). Entities can be classified into two groups, i.e., proper nouns and common nouns. Proper nouns are also known as Named Entities (NE) and correspond to the name of people, organizations or locations, e.g., *John*, *WHO* or *Canada*. Common nouns describe classes of objects, e.g., *spoon* or *cancer*.

One possible solution to address the remaining issues of MT regarding semantic ambiguity and OOV words lies in the use of KGs, which have emerged over recent decades as a paradigm to make the semantics of data explicit so that it can be used by machines (Berners-Lee et al., 2001). KGs are a family of flexible knowledge representation paradigm intended to facilitate the processing of knowledge for both humans and machines. KGs (especially KGs in the RDF format) commonly stores knowledge in triples. Each triple consists of

1. a subject which is often an entity.
2. a relation which is often called a property.
3. an object which is an entity or a literal.[4]

For example, the following triple expresses that Albert Einstein was born in Ulm:

```
:Albert_Einstein :birthPlace :Ulm .
```

[4] a string or a value with a unit

The explicit semantic knowledge in KGs can enable MT systems to supply translations with significantly better quality while maintaining the translation process scalable (Heuss, 2013). In addition, the disambiguated knowledge about real-world entities, their properties, and relationships can potentially be used to infer the right meaning of ambiguous sentences or words as well as improve the performance of MT systems on the reordering task.

Recent work has successfully exploited the apparent opportunity of using KGs for the improvements of other NLP tasks such as NLI (K M et al., 2018), QA (Sorokin and Gurevych, 2018), and Machine Reading (MR) (Yang and Mitchell, 2017). According to Moussallem et al. (2018c), the distinct opportunity of using KGs for MT has already been studied by several approaches. However, none had defacto implemented and used the benefits of KGs in the training phase of NMT before this work.

NLG is the task of automatically converting non-linguistic data into coherent natural language text (Reiter and Dale, 2000; Gatt and Krahmer, 2018). Recently, a new line of research has emerged, which relies on KGs as input data. It has a task named RDF-to-Text, which generates texts from RDF KGs (Colin et al., 2016). This task is an extension of MT as understood classically given that it translates from a non-natural to a natural language. Therefore, we envisage that it will help enhance the fluency in language translation.

In this thesis, we devise novel approaches that rely on KGs to improve the disambiguation, translation, and generation of entities in texts. Section 1.1 specifies the problems and identifies motivation and research challenges. Section 1.2 summarizes the contributions of the thesis.

1.1 PROBLEM SPECIFICATION AND CHALLENGES

A large number of MT approaches have been developed over the last two decades. For instance, translators began by using methodologies based on linguistics, which led to the family of Rule-Based Machine Translation (RBMT)(Arnold, 1994). However, RBMT systems have a critical drawback in their reliance on manually crafted rules, thus making the development of new translation modules for different languages even more difficult as each language has its own syntax (Costa-Jussa et al., 2012; Thurmair, 2004). SMT and Example-Based Machine Translation (EBMT) were developed to deal with the scalability issue in RBMT (Brown et al., 1990), a necessary characteristic of MT systems

that handle data at Web scale. Presently, these approaches have begun to address the drawbacks of rule-based approaches. However, certain problems that had already been solved for RBMT methods reappeared. The majority of these problems are connected to the issue of ambiguity, including syntactic and semantic variations (Koehn, 2010). Subsequently, RBMT and SMT have been combined to resolve the drawbacks of these two families of approaches. This combination of methods is called hybrid MT. Although hybrid approaches have been achieving good results, they still suffer from some of the limitations of RBMT (Costa-Jussa and Fonollosa, 2015; Costa-jussà, 2015; Thurmair, 2009). For example, the creation of manually crafted rules to handle syntax divergences.

Below, we detail some key MT challenges, which were unresolved when we began our work and still experienced by the MT approaches aforementioned (Moussallem et al., 2018c):

1. *Complex semantic ambiguity*: This challenge is mostly caused by the existence of homonyms, polysemous words, and named entities. Homonyms are different words that mean different things but share the same orthographic and phonological forms. For example, "`bank`" can mean "the land alongside or sloping down to a river or lake" or "financial organization". Polysemous words are considered as the same word but with different, still related senses. For instance, "`wood`" can refer to a piece of a tree or a collection of many trees. MT systems commonly struggle to translate these words correctly, even if the models are built upon n-grams with large *n* (e.g., 7-grams). Therefore, a significant amount of parallel data is usually necessary to translate such words and expressions adequately. However, data is not only the main aspect to consider while learning translations. For example, some homonyms such as "`kiwi`" can also refer to a named entity, and therefore it requires more specific learning features than a vast amount of parallel data to determine its correct meaning.

2. *Structural divergence*: By definition, structural reordering is reorganizing the order of the syntactic constituents of a language according to its original structure (Bisazza and Federico, 2016). It, in turn, is a critical issue because fluency is one of the key aspects in the translation process. Every language has its own syntax. Thus an MT system, which aims to translate a given language pair, needs to have an adequate model for the syntax of the involved languages. For instance, reordering a sentence from Japanese to English is one of the most challenging

techniques because of the SVO (subject-verb-object) and SOV (subject-object-verb) word-order difference. One English word often groups multiple meanings of Japanese characters. For example, Kanji (Japanese) characters make subtle distinctions between homonyms that would not be clear in a phonetic language such as English. The following words, 史 (history), 師 (teacher), 市 (a market or city), 矢 (arrow), 士 (a warrior or gentleman) are pronounced as (shi), the same as “she" (English feminine pronoun).

3. *Linguistic properties/features*: A large number of languages display a complex tense system. When confronted with sentences from such languages, it can be hard for MT systems to recognize the current input tense and to translate the input sentence into the right tense in the target language. For instance, some irregular verbs in English like “set” and “put” cannot be determined to be in the present or past tense without previous knowledge or pre-processing techniques when translated to morphologically rich languages, e.g., Portuguese, German or Slavic languages. Additionally, the grammatical gender of words in such morphologically rich languages contributes to the problem of tense generation where a certain MT system has to decide which inflection to use for a given word. This challenge is a direct consequence of the structural reordering issue and remains a significant problem for modern translator systems.

Additionally, recent literature suggests 5 different challenges, which are described more generically below (Lopez and Post, 2013).

1. Recent work focuses excessively on English and European languages as one of the involved languages in MT approaches. In addition, there is a lack of research on low-resource language pairs such as African and/or South American languages.

2. Previous work shows limitations of SMT approaches for translating across domains. Most MT systems exhibit good performance on legislative domains due to a large amount of data provided by the European Union. In contrast, translations performed on sports and life-hacks commonly fail because of the lack of training data.

3. Few MT approaches are able to translate non-standard speech texts from social networks (e.g., tweets). This kind of text poses several challenges for MT systems, such as syntactic variations.

4. There is a shortage of MT approaches for translating morphologically rich languages. This challenge shares the same problem with the first one, namely the excessive focus on English as one of the involved languages. Therefore, MT systems that translate content between, for instance, Arabic and Spanish, are rare.

5. For the speech translation task, the bilingual parallel data, which are used for training the MT models, differs widely from real user speech.

The challenges above are clearly not independent, which means that addressing one of them can have an impact on the others. We focus on the portions of these problems related to entities. Entities are found in a KG, where they are described within triples (Auer et al., 2007; Vrandečić and Krötzsch, 2014). It is already clear that the real benefit of KGs comes from their capacity to provide unseen knowledge about emergent data, which appears every day. Thus, our central research question can be stated as follows:

RQ. Can KGs alleviate the ambiguity problem and be used to improve the quality of automatic text translation and generation?

In the following subsections, we present the challenges that need to be tackled to answer our central research question.

1.1.1 CHALLENGE 1: MULTILINGUAL ENTITY DISAMBIGUATION

Understanding the EL task in a multilingual environment is the first step to discern how to deal with entities in text translation and generation. One of the most important MT tasks is EL, also known as Named Entity Disambiguation (NED). The goal of EL is the disambiguation of entities and common words (concepts and terminologies) in texts. Disambiguation refers to the process of removing the ambiguity of words by identifying their single semantic meaning for a particular context, in our case entities. Formally, the goal of EL algorithm is as follows: given a piece of text, a reference knowledge base K, and a set of entity mentions in that text, map each entity mention to the corresponding resource in K. Several challenges have to be addressed when implementing an EL system. First, an entity can have a large number of surface forms (SF) (also known as labels) due to synonymy, acronyms, and typos. For example, `New York City`, `NY` and `Big Apple` are labels for the same entity. Moreover, multiple entities can share the same

name due to homonymy and ambiguity. For example, both the state and the city of New York are called `New York`.

Despite the complexity of the task, EL approaches have recently been achieving increasingly better results by relying on trained machine learning models (Röder et al., 2018). A portion of these approaches claim to be multilingual, and most of them rely on models that are trained on English corpora with cross-lingual dictionaries. However, these underlying models being trained on English corpora make them prone to errors when migrated to a different language. Additionally, such approaches rarely make their models or data available on more than three languages due to the lack of training data (Röder et al., 2018).

A large number of multilingual approaches have been developed over recent years-(Ganea et al., 2016). However, to the best of our knowledge, no work has investigated the real disambiguation capability of KGs in a broader multilingual and deterministic context. Thus, our first goal is to investigate the disambiguation task based on KGs and analyze whether they can contribute to the translation of entities. Hence, we derive the following research questions:

RQ1. Can a KG-based EL approach achieve a similar F-score performance across languages?

RQ2. Does a language-based KG influence the disambiguation quality of entities in multilingual sentences?

1.1.2 CHALLENGE 2: TEXT GENERATION WITH ENTITIES

The input data in RDF-to-Text consists of entities and the relations between them, therefore generating references for these entities is a core task in many NLG systems (Krahmer and Van Deemter, 2012a). REG, the task responsible for generating these references, is typically presented as a two-step procedure. First, the referential form needs to be chosen, asking whether a reference at a given point in the text should assume the form of, for example, a proper noun (“Stephen Hawking”), a pronoun (“he/him/his”) or description (“the physicist”). Second, the REG model must account for the different ways in which a particular referential form can be realized. For example, both “Stephen” and “Hawking” are name variants of Stephan Hawking that may occur in a text. He can also alternatively be described as, say, “the brilliant scientist”.

A generic NLG pipeline is composed of three tasks - *document planing*, *micro planning* and *realization*. Before generating the respective referring expressions for the entities, several steps have to be taken into account. For example, Listing 1.1 shows a fragment of Stephen Hawking sub-KG[5] which represents the following information: *"Stephen Hawking was a scientist who worked in physics. He was born in Oxford and died in Cambridge."*.

```
:Stephen_Hawking :type :Scientist
:Stephen_Hawking :deathPlace :Cambridge
:Stephen_Hawking :field :Physics
:Stephen_Hawking :birthPlace :Oxford
```

Listing 1.1: An excerpt of RDF triples.

Even though the generation of natural language from KGs has gained substantial attention (Colin et al., 2016), English is the only language that has been widely targeted. Only a few authors (e.g., Keet and Khumalo (2017) for IsiZhulu) have exploited the generation of other languages. Consequently, there is a lack of multilingual approaches for the generation of texts from RDF KGs. Additionally, most of the earlier REG approaches focus either on selecting referential forms (Orita et al., 2015; Castro Ferreira et al., 2016), or on selecting referential content, typically zooming in on one specific kind of reference such as pronouns (Henschel et al., 2000; Callaway and Lester, 2002), definite descriptions (Dale and Haddock, 1991), or proper noun generations (Siddharthan et al., 2011; van Deemter, 2016; Castro Ferreira et al., 2017). Therefore, no previous work has addressed the full REG task, which given a number of entities in a text, produces corresponding referring expressions by simultaneously selecting both form and content. Moreover, in previous models, notions such as *salience* play a central role, where it is assumed that entities, which are salient in the discourse, are more likely to be referred to using shorter referring expressions (like a pronoun) than less salient entities, which are typically referred to using longer expressions (like full proper nouns).

Although some basic linguistics mistakes have been solved by Neural Network-based approaches, the lack of complex models for linguistic rules still causes ambiguity problems in text generation (e.g., errors on relative pronouns) (Bisazza and Federico, 2016). The issues mentioned above leads to the following research question:

[5]http://dbpedia.org/resource/Stephen_Hawking

RQ3: Can KGs as input support the generation of multilingual text?

RQ4: Can KGs be used for accomplishing the full REG task?

1.1.3 CHALLENGE 3: ENTITY TRANSLATION IN TEXTS

Entities are a common and arduous problem across different NLP tasks. Regarding MT, Named Entity (NE)'s primary issue is caused by common words from a source language that are used as proper nouns in a target language. For instance, the word "Kiwi" is a family name in New Zealand which comes from the Māori culture, but it also can be a fruit, a bird, or a computer program. Most words have multiple interpretations depending on the context in which they are mentioned. In the MT field, Word Sense Disambiguation (WSD) techniques involve finding the respective meaning and correct translation to these ambiguous words in target languages. This ambiguity problem was identified early in MT development. In 1960, Bar-Hillel (1960) stated that an MT system is not able to find the right meaning without specific knowledge. Although the ambiguity problem has been lessened significantly since the contribution of Carpuat and subsequent works (Carpuat and Wu, 2007; Navigli, 2009; Costa-Jussà and Farrús, 2014), this problem remains a challenge.

According to Moussallem et al. (2018c), KGs were applied mainly to the output translation of Phrase-Based Statistical Machine Translation (PBSMT) approaches in the target language as a post-editing technique. Although applying this technique has increased the quality of a translation, it is tedious to implement when common words have to be translated instead of named entities, then be applied several times to achieve a successful translation. In MT systems, dealing with entities is directly related to the ambiguity problem. Therefore, we argue that the entity problem has to be resolved in that broader context.

Recently, NMT models have shown significant improvements in translation and have been widely adopted given their sustained improvements over the previous state-of-the-art PBSMT approaches (Koehn et al., 2007). A number of NN architectures have therefore been proposed in the recent years, ranging from recurrent (Bahdanau et al., 2014; Sutskever et al., 2014) to self-attentional networks (Vaswani et al., 2017). A given NMT model is basically trained to maximize the likelihood of each token in the target sentence, by taking into account the source sentence and the previous target tokens as

input. However, a major drawback of NMT models is that they need large amounts of training data to return adequate results and have a limited vocabulary size due to their computational complexity (Luong and Manning, 2016). The data sparsity problem in MT, which is mostly caused by a lack of training data, manifests itself particularly in the poor translation of rare and OOV words, e.g., entities or terminological expressions rarely or never seen in the training phase.

Previous work has attempted to deal with entities and the data scarcity by introducing character-based models (Luong and Manning, 2016) or BPE algorithms (Sennrich et al., 2016a). Additionally, different strategies were developed for overcoming the lack of training data, such as back-translation (Sennrich et al., 2016b), which relies on the use of monolingual data being translated by a different NMT model and added as additional synthetic training data. Moreover, the benefits of incorporating type information on entities—e.g., NE-tags such as `PERSON`, `LOCATION` or `ORGANIZATION`—into NMT by relying on Named Entity Recognition (NER) systems have been shown in previous works (Ugawa et al., 2018; Li et al., 2018). Despite the significant advancement of previous work in NMT, translating entities and terminological expressions remains a challenge (Koehn and Knowles, 2017) and none of the above mentioned approaches have exploited the application of KGs in NMT systems. Hence it leads to the following research question:

RQ5: Can an NMT model enhanced with a bilingual KG improve translation quality?

1.1.4 CHALLENGE 4: LOW-RESOURCE KNOWLEDGE GRAPHS

Considerable amounts of partly human effort have been invested in making KGs available across languages. However, even popular KGs like DBpedia and Wikidata are most abundant in their English version (Lakshen et al., 2018). Additionally, region-specific facts are often limited to the KG specific to the region from which they emanate or to the KG in the language spoken in said region (Aprosio et al., 2013). This lack of multilingual knowledge availability limits the porting of NLP tasks such as EL, NLG, and NMT to different languages.

Previous works have tried to address the translation of KGs by carrying out a localization task that relies on SMT systems for translating the labels of KGs into target languages. This kind of approach ignores an essential part of a KG, namely its graph structure. For example, considering a highly ambiguous label in DBpedia KG such

as *Kiwi*, an MT system has to predict in which sub-KG domain this word has to be translated in the target language. Otherwise, *Kiwi* can be erroneously translated to the common term for inhabitants of New Zealand,[6] or a bird,[7] thus affecting the structure and alignment quality of the translated KG. These domains can be derived in KGs through predicates such as type predicates (i.e., `rdf:type` in RDF). Taking the graph structure of KG into account can support an MT system when spotting the correct translation for ambiguous labels. Few works have designed approaches to tackle this problem. Hence we investigate the following research questions:

RQ6: Can NMT support a full (Uniform Resource Identifier (URI)s and labels) translation of KGs?

RQ7: Can an artificially enriched KG improve the performance of a system on NLP tasks?

1.2 THESIS OVERVIEW

1.2.1 CONTRIBUTIONS

In the following, our contributions are summarized according to each of the challenges aforementioned.

Challenge 1: Multilingual Entity Disambiguation

Contribution 1: This drawback is addressed by presenting a novel multilingual, knowledge-base agnostic and deterministic approach to entity linking, dubbed MAG. MAG is based on a combination of context-based retrieval on structured knowledge bases and graph algorithms. We evaluate MAG on 23 datasets and in 7 languages. Our results show that MAG achieves state-of-the-art performance on English datasets and outperforms all other approaches on non-English languages (Moussallem et al., 2017). Further, we extend MAG to 40 languages and deploy two versions as demos - one using DBpedia, another Wikidata (Moussallem et al., 2018b). The demos answer on average more than 170,000 requests per year.

[6] http://dbpedia.org/resource/Kiwi_(people)
[7] http://dbpedia.org/resource/Kiwi

Challenge 2: Text Generation with Entities

Contribution 2: We address this research gap by presenting RDF2PT, an approach that verbalizes RDF data to Brazilian Portuguese. We evaluate RDF2PT in an open questionnaire, with 44 native speakers divided into experts and non-experts. Our results suggest that RDF2PT is able to generate text similar to that generated by humans and can hence be easily understood (Moussallem et al., 2018a). Afterward, we extend RDF2PT to Spanish (Ngonga Ngomo et al., 2018) and English (Ngonga Ngomo et al., 2019).

Contribution 3: Traditionally, REG models first decide on the form and then on the content of references to discourse entities in text and rely thereby on features such as salience and grammatical function. No previous work has investigated either how to tackle both sub-tasks at once or use RDF KG as input to this task. We handle this problem by presenting the first approach relying on deep neural networks, which makes decisions about form and content in one go without explicit feature extraction. Using RDF KG (Bonatti et al., 2019), the neural model substantially improves over two strong baselines (Ferreira et al., 2018b). We also extend our training data to the German language, making it able to generate referring expressions in German (Ferreira et al., 2018a).

Challenge 3: Entity Translation in Texts

Contribution 4: While neural networks have led to substantial progress in machine translation, their success depends heavily on large amounts of training data. However, parallel training corpora are not always readily available. Out-of-vocabulary words, mostly entities and terminological expressions, pose a difficult challenge to NMT systems. We alleviate this problem by implementing the first KG-augmented NMT model, named KG-NMT. We use knowledge graph embeddings to enhance the semantic feature extraction of neural models. Thus, this approach optimize the translation of entities and terminological expressions in texts, consequently leading to better translation quality. Our knowledge-graph-augmented neural translation model, dubbed *KG-NMT*, achieve significant and consistent improvements of +3 BLEU, METEOR and CHRF3 on average on the *newstest* datasets between 2015 and 2018 for the WMT English-German translation task (Moussallem et al., 2019a).

Challenge 4: Low-resource Knowledge Graphs

Contribution 5: We address the current limitations of knowledge graphs w.r.t. multilinguality by proposing THOTH, the first full neural-based approach for translating and enriching knowledge graphs across languages. THOTH extracts bilingual alignments between a source and target knowledge graph and learns how to translate from one to the other by relying on two different recurrent neural network models along with knowledge graph embeddings. We evaluate THOTH extrinsically by comparing the German DBpedia with the German translation of the English DBpedia on two tasks: fact checking and entity linking. In addition, we run a manual intrinsic evaluation of the translation. Our results show that THOTH is a promising approach since it achieves a translation accuracy of 88.56%. Moreover, its enrichment improves the quality of the German DBpedia significantly, as we report +18.4% accuracy for fact validation and +19% F_1 for entity linking (Moussallem et al., 2019b).

The main contributions of this thesis can be summarized as follows:

1. A multilingual, knowledge-base agnostic and deterministic entity-linking approach for 40 languages (Moussallem et al., 2017, 2018b)
2. A multilingual RDF-to-text approach that works on Brazilian Portuguese, Spanish and English and is virtually extensible for German, French and Italian (Moussallem et al., 2018a)
3. The first NN model for tackling the full REG task by using KGs (Ferreira et al., 2018b)
4. The first NMT model augmented with KGs (Moussallem et al., 2019a)
5. The first full KG translation and enrichment approach based on NN models (Moussallem et al., 2019b)

1.2.2 STRUCTURE

Having motivated our work in this chapter, Chapter 2 summarizes the main contributions of the thesis and presents them in a coherent way. Chapter 3 concludes the thesis with

a summary and outlook on future research directions. The publications underlying this cumulative thesis can be found in Appendix A along with a detailed breakdown of the contributions of individual authors in Appendix B.

2
Contributions

This chapter describes the main contributions of this thesis: in Section 2.1, we present a multilingual EL approach; in Section 2.2, we unveil a Brazilian Portuguese RDF-KG-based NLG approach; in Section 2.3, we discuss a neural-based REG model; in Section 2.4, we give some insights into a KG-augmented NMT model; in Section 2.5, we develop a neural-based approach for translating and enriching KGs.

2.1 MAG: A Multilingual, Knowledge-base Agnostic and Deterministic Entity Linking Approach

FOR ALLEVIATING THE LACK OF MULTILINGUAL EL APPROACHES 1.1.1 , we devised a multilingual, knowledge-base agnostic and deterministic approach, named MAG. The EL process implemented by MAG consists of two phases. Several indexes are generated during the offline phase. The entity linking per se is carried out during the online phase and consists of two steps: 1) candidate generation and 2) disambiguation. An overview can be found in Figure 2.1.

2.1.1 Offline Index Creation

MAG relies on the following five indexes: surface forms, person names, rare references, acronyms and context.

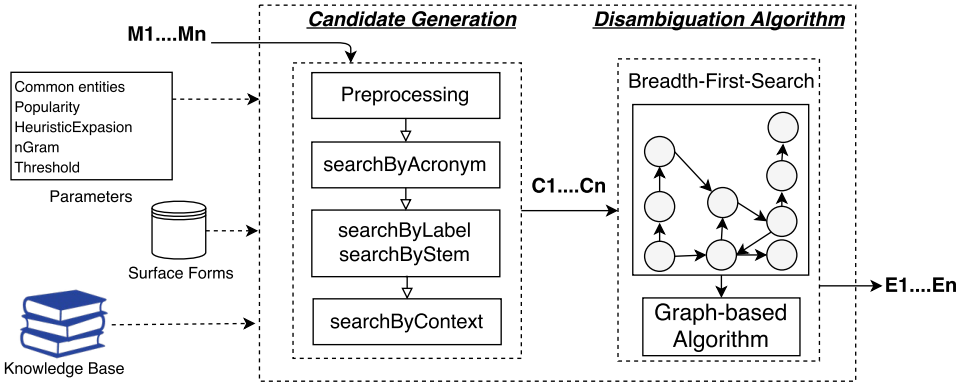

Figure 2.1: MAG architecture overview.

Surface forms. MAG relies exclusively on structured data to generate surface forms for entities so as to remain KB-agnostic. For each entity in the reference Knowledge Base (KB), our approach harvests all labels of the said entity as well as its type and indexes them.Additional SFs can be collected from different sources (Usbeck et al., 2014; Bryl et al., 2016).

Person names - This index accounts for the variations in names for referencing persons (Krahmer and Van Deemter, 2012b) across languages and domains. Persons are referred to by different portions of their names. For example, the artist `Beyoncé Giselle Knowles-Carter` is often referred to as `Beyoncé` or `Beyoncé Knowles`. Moreover, languages such as Chinese and Japanese put the family name in front of the given name (in contrast to English, where names are written in the reverse order). Our technique handles the problem of labelling persons by generating all possible permutations of the words within the known labels of persons and adding them to the index of names.

Rare references - This index is created if textual descriptions are available for the resources of interest (e.g., if resources have a `rdfs:comment` property). A large number of textual entity descriptions provide type information pertaining to the resource at hand, as in the example "Michael Joseph Jackson was an American singer ..."[1]. Hence, we use different language versions of the Stanford POS tagger (Toutanova and Manning, 2000) and TreeTagger[2] on the first line of a resource's description and collect any noun

[1] See `rdfs:comment` of http://dbpedia.org/resource/Michael_Jackson.
[2] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

phrase that contains an adjective. For example, we can extract the supplementary SF `American singer` for our example.

Acronyms - Acronyms are used across a large number of domains, e.g., in news (see AIDA and MSNBC datasets). We thus reuse a handcrafted index from STANDS4.[3]

Context - Our context index relies on the Concise Bounded Description (CBD)[4] of resources. The literals found in the CBD of each resource are first freed of stop words. Then, each preprocessed string is added as an entry that maps to the said resource.

2.1.2 CANDIDATE GENERATION

The candidate generation and the disambiguation steps occur online, i.e., when MAG is given a document and a set of mentions to disambiguate. The goal of the candidate generation step is to retrieve a tractable number of candidates for each of the mentions. These candidates are later inserted into the disambiguation graph, which is used to determine the mapping between entities and mentions (see Section 2.1.3).

First, we **preprocess mentions** to improve the retrieval quality using well-known pre-processing NLP techniques such as regular expressions, lemmatization, stemming and true casing.

The second step of the candidate generation, the **candidate search**, is divided into three parts:

By Acronym - If a mention is considered an acronym by our preprocessing, we expand the mention with the list of possible names from the acronym index mentioned above. For example, “PSG" is replaced by “Paris Saint-Germain".

By Label - First, MAG retrieves candidates for a mention using exact matches to their respective principal reference. For example, the mention “Barack Obama" and the principal reference of the former president of the USA, which is also “Barack Obama", match exactly. In cases it finds a string similarity match with the main reference of 1.0, the remaining steps are skipped. If this search does not return any candidates, MAG starts a new search using a trigram similarity threshold σ over the SF index. In cases where the set of candidates is still empty, MAG stems the mention and repeats the search. For example, MAG stems “Northern India" to “North India" to account for linguistic variability.

[3] See http://www.abbreviations.com/
[4] https://www.w3.org/Submission/CBD/

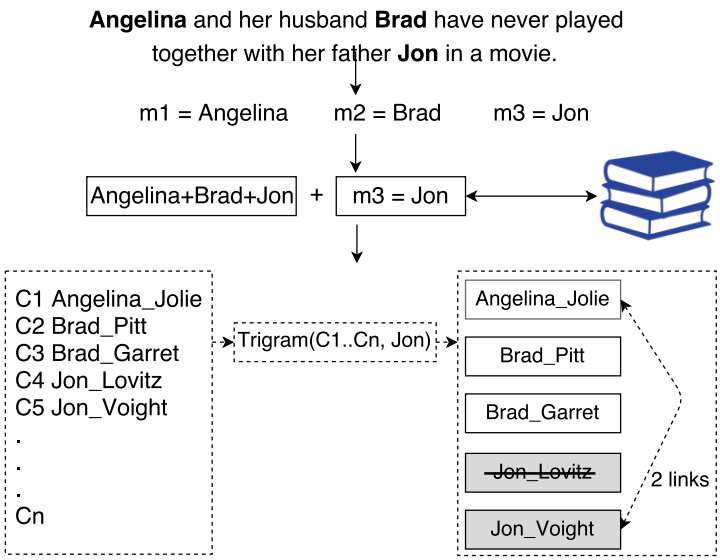

Figure 2.2: Search using the context index. White boxes on the right side depict candidates discarded by the trigram filter.

By Context - Here, two **post-search filters** are applied to find possible candidates from the context index. Before applying both filters, MAG extracts all entities contained in the input document. These entities are used as an addition while searching a mention in the context index. This search relies on TF-IDF (Ramos et al., 2003) which reflects the importance of a word or string in a document corpus relative to the relevance in its index. Afterwards, MAG first filters unlikely candidates by applying trigram similarity. Second, MAG retrieves all direct links among the remaining candidates in the KB. Our approach uses the number of connections to find highly related entity sets for a specific mention. This is similar to finding a dense subgraph (Hoffart et al., 2011). Figure 2.2 illustrates an example which contains three ambiguous entities, namely “Angelina", “Brad" and “Jon". Regarding the mention “Jon", MAG searches the context index using “[(Angelina + Brad + Jon) + Jon]" as a query. MAG keeps only “Jon_Lovitz" and “Jon_Voight" after trigram filtering. Only “Jon_Voight", the father of “Angelina_Jolie", has direct connections with the other candidates and is thus chosen.

To improve the quality of candidates, ranking entities according to their popularity is an appropriate factor. If MAG makes use of this ranking configuration, the number of candidates retrieved from the index is increased and then the result is sorted. Afterward, MAG returns the top 100 candidates. The popularity is calculated using Page Rank (Page

et al., 1999) over the underlying KB.In case it is unable to leverage Page Rank on certain KB, it falls back to a heuristic of inlinks and outlinks.

2.1.3 ENTITY DISAMBIGUATION ALGORITHM

After the candidate generation step, the computation of the optimal candidate to mention assignment starts by constructing a disambiguation graph G_d with depth d similar to the approach of AGDISTIS.

Definition 2.1 *Knowledge Base: We define KB K as a directed graph $G_K = (V, E)$ where the nodes V are resources of K, the edges E are properties of K and $x, y \in V, (x, y) \in E \Leftrightarrow \exists p : (x, p, y)$ is a triple in K.*

Given the set of candidates C, we begin by building an initial graph $G_0 = (V_0, E_0)$ where V_0 is the set of all resources in C and $E_0 = \emptyset$. Starting with G_0 the algorithm expands the graph using Breadth-First-Search (BFS) technique in order to find hidden paths among candidates. The extension of a graph is $G_i = (V_i, E_i)$ to a graph $\rho(G_i) = G_{i+1} = (V_{i+1}, E_{i+1})$ with $i = 0, \ldots, d$. The ρ (BFS) operator iterates d times on the input graph G_0 to compute the initial disambiguation graph G_d. After constructing G_d, it needs to identify the correct candidate node for a given mention. Here, we rely on HITS (Kleinberg, 1999) or Page Rank (Page et al., 1999) as disambiguation graph algorithms. This choice comes from a comparative study of the differences between both (Devi et al., 2014).

HITS uses hub and authority scores to define a recursive relationship between nodes. An authority node is a node that many hubs link to and a hub is a node that links to many authorities. The authority values are equal to the sum of the hub scores of each node that points to it. The hub values are equal to the sum of the authority scores of each node that it points to. According to previous work (Usbeck et al., 2014), we chose 20 iterations for HITS which suffice to achieve convergence in general.

Page Rank has a wide range of implementations. We implemented the general version in accordance with (Page et al., 1999). Thus, we defined the possibility of jumping from any node to any other node in the graph during the random walk with a probability $\alpha = (1 - w) = 0.15$. We empirically chose 50 Page Rank iterations which has shown to be a reasonable number for EL (Zwicklbauer et al., 2016). We assigned a standard weight $w = 0.85$ for each node. Finally, the sum is calculated by spreading the current weight divided by outgoing edges.

Independent of the chosen graph algorithm, the highest candidate score among the set of candidates C is chosen as correct disambiguation for a given mention m_i. Note, MAG also considers emergent entities (Hoffart et al., 2014) and assigns a new URI to them.[5]

2.1.4 EVALUATION

We measured the performance of MAG on 17 datasets and compared it to the state of the art for EL in English. Second, we evaluated MAG's portability to other languages. To this end, we compared MAG and the multilingual state of the art using 6 datasets from different languages. For both evaluations we use HITS and Page Rank. Third, we carried out a fine-grained evaluation providing a deep analysis of MAG using the method proposed in (Waitelonis et al., 2016). Throughout our experiments, we used DBpedia as reference KB. For our overall evaluation, we relied on the GERBIL platform (Usbeck et al., 2015) and integrated all datasets into it for the sake of comparability.

2.1.5 RESULTS

On English datasets. The English results are shown in the first part of Table 2.1. An analysis of our results shows that although the acronym index is an interesting addition for potential improvements, its contribution amounts only to 0.05% F-measure on average over all datasets. Also, the popularity feature improves the results in almost every data set. It can be explained by the analysis of (Waitelonis et al., 2016), which demonstrates that most datasets were created using more popular entities as mentions. Thus, this bias eases their retrieval[6]. HITS has shown better results on average than Page Rank.[7] However, Page Rank did show promising results in some datasets (e.g., Spotlight corpus, AQUAINT, and N3-RSS-500). MAG using HITS outperformed the other approaches on 4 of the 17 datasets while achieving comparable results on others, e.g., ACE2004, MSNBC, and OKE datasets.

On Multilingual datasets. Here, we show the easy portability and high quality of MAG for many different languages. Next to German, Italian, Spanish, French and Dutch, we chose Japanese to show the promising potential of MAG across different language systems. MAG's preprocessing NLP techniques are multilingual, thus there is

[5] https://www.w3.org/TR/cooluris/
[6] see the results without popularity using HITS http://gerbil.aksw.org/gerbil/experiment?id=201701220014
[7] http://gerbil.aksw.org/gerbil/experiment?id=201701240030

Table 2.1: Micro F-measure across approaches. Red entries are the top scores while blue represents the second best scores.

Language	**Tools/datasets**	***AGDISTS***	***AIDA***	***Babelfy***	***DBpedia***	***DoSer***	***entityclassifier.eu***	***FRED***	***Kea***	***NERD-ML***	***PBOH***	***WAT***	***xLisa***	***MAG + HITS***	***MAG + PR***
English	ACE2004	0.65	0.70	0.53	0.48	**0.75**	0.50	0.00	0.66	0.58	**0.72**	0.66	0.70	0.69	0.60
	AIDA/CoNLL-Complete	0.55	0.68	0.66	0.50	0.69	0.50	0.00	0.61	0.20	**0.75**	**0.71**	0.48	0.59	0.54
	AIDA/CoNLL-Test A	0.54	0.67	0.65	0.48	0.69	0.48	0.00	0.61	0.00	**0.75**	**0.7**	0.45	0.59	0.54
	AIDA/CoNLL-Test B	0.52	0.69	0.68	0.52	0.69	0.48	0.00	0.61	0.00	**0.75**	**0.72**	0.47	0.57	0.52
	AIDA/CoNLL-Training	0.55	0.69	0.66	0.50	0.69	0.52	0.00	0.61	0.28	**0.75**	**0.71**	0.48	0.60	0.55
	AQUAINT	0.52	0.55	0.68	0.53	**0.82**	0.41	0.00	0.78	0.60	**0.81**	0.73	0.76	0.67	0.68
	Spotlight	0.27	0.25	0.52	0.71	**0.81**	0.25	0.04	0.74	0.56	**0.79**	0.67	0.71	0.65	0.66
	IITB	0.47	0.18	0.37	0.30	0.43	0.14	0.00	**0.48**	0.43	0.38	0.41	0.27	**0.52**	0.43
	KORE50	0.27	**0.70**	**0.74**	0.46	0.52	0.30	0.06	0.60	0.31	0.63	0.62	0.51	0.24	0.24
	MSNBC	0.73	0.69	0.71	0.42	**0.83**	0.51	0.00	0.78	0.62	**0.82**	0.73	0.5	0.79	0.75
	Microposts2014-Test	0.33	0.42	0.48	0.50	**0.76**	0.41	0.05	0.64	0.52	**0.73**	0.60	0.55	0.45	0.44
	Microposts2014-Train	0.42	0.51	0.51	0.48	**0.77**	0.00	0.31	0.65	0.52	**0.71**	0.63	0.59	0.49	0.44
	N3-RSS-500	0.66	0.45	0.44	0.20	0.48	0.00	0.00	0.44	0.38	0.53	0.44	0.45	**0.69**	**0.67**
	N3-Reuters-128	0.61	0.47	0.45	0.33	**0.69**	0.00	0.41	0.51	0.41	**0.65**	0.52	0.39	**0.69**	0.64
	OKE 2015 Task 1 evaluation	0.59	0.56	0.59	0.31	0.59	0.00	0.46	**0.63**	0.61	**0.63**	0.57	**0.62**	0.58	0.55
	OKE 2015 Task 1 example	0.50	**0.60**	0.40	0.22	0.55	0.00	**0.60**	0.55	0.00	0.50	**0.60**	0.50	**0.67**	0.50
	OKE 2015 Task 1 training	0.62	0.67	0.71	0.25	**0.78**	0.00	0.61	**0.78**	**0.77**	0.76	0.72	0.75	0.72	0.70
Multilingual	N^3 news.de	0.61	0.52	0.50	0.48	0.56	0.28	0.00	0.61	0.33	0.30	0.59	0.36	**0.76**	**0.63**
	Italian Abstracts	0.22	0.28	**0.33**	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	**0.80**	**0.80**
	Spanish Abstracts	0.25	0.33	0.26	0.00	0.24	0.27	0.00	0.47	0.00	0.31	0.33	0.31	**0.75**	**0.68**
	Japanese Abstracts	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	**0.38**	0.00	0.00	**0.54**	**0.54**
	Dutch Abstracts	0.33	0.36	0.36	0.28	0.36	0.22	0.00	0.40	0.00	0.5	0.40	0.25	**0.66**	**0.67**
	French Abstracts	0.00	0.00	**0.28**	0.22	0.00	0.25	0.00	0.00	0.00	0.20	**0.28**	**0.28**	**0.80**	**0.80**
	Average	0.45	0.48	0.50	0.36	0.55	0.24	0.11	0.53	0.31	**0.59**	0.54	0.45	**0.63**	**0.59**
	Standard Deviation	0.19	0.22	0.18	0.19	0.27	0.21	0.21	0.23	0.26	0.20	0.22	0.20	0.13	0.13

no additional implementation for handling the mentions with different characters. We used the same set of parameters as in the English evaluation but excluded the acronyms as they were only collected for English. Moreover, we performed the Page Rank algorithm over each KB in each respective language to collect popularity values of their entities. The results displayed in the second part of Table 2.1 show that MAG, using HITS, outperform all publicly available state-of-the-art approaches. Additionally, for Dutch, Page Rank outperforms HITS score. The improved performance of MAG is due to its knowledge-base agnostic algorithms and indexing models. For instance, although the mention “Obama" has a high popularity in English, it may have less popularity

in Italian or Spanish KBs. Studies about the generation of proper names support this observation (Dale and Reiter, 1995; Ferreira et al., 2017).

Fine-Grained Evaluation. In this analysis, we measured the quality of a given EL for linking different types of entities. This extension also considers that a corpus tends to focus strongly on prominent or popular entities, which may cause evaluation problems. Hence, the extension evaluates the capability of a given EL system to find entities with different levels of popularity, thus revealing its degree of bias towards popular entities. The fine-grained analysis shows that MAG is better at linking persons than other types of entities, and it can be explained by the indexes created by MAG in the offline phase. They collect last names and rare surfaces for entities. In addition, the results show that MAG is not biased towards linking only popular entities, as can be seen in Table 2.2.

Table 2.2: Fine-grained micro F1 evaluation.

Filter	**IITB**	**N3-RSS-500**	**MSNBC**	**Spotlight**	**N3-Reuters-128**	**OKE 2015**
Persons	0.95	0.83	0.94	0.84	0.80	0.92
Page Rank 10%	0.73	0.67	0.83	0.74	0.79	0.76
Page Rank 10%-55%	0.72	0.72	0.70	0.69	0.73	0.79
Page Rank 55%-100%	0.73	0.71	0.73	0.75	0.76	0.82
Hitsscore 10%	0.77	0.74	0.77	0.69	0.73	0.76
Hitsscore 10%-55%	0.69	0.66	0.64	0.69	0.79	0.78
Hitsscore 55%-100%	0.71	0.66	0.84	0.74	0.77	0.80

2.1.6 DEMONSTRATION

This demonstration extends MAG to support EL in 40 different languages, including especially low-resources languages such as Ukrainian, Greek, Hungarian, Croatian, Portuguese, Japanese and Korean. Our demo relies on online web services which allow for an easy access to our entity linking approaches and can disambiguate against DBpedia and Wikidata. Additionally, MAG supports POST requests as well as it has a user-friendly web interface.

Figure 2.3: A screenshot of MAG's web-based demo working on Spanish.

2.1.7 REPRODUCIBILITY

MAG was implemented within the AGDISTIS framework.[8] In addition, all experimental data, code and, results are publicly available.[9]

2.1.8 SUMMARY

The main contributions of this paper can be summarized as follows:

- We present a novel multilingual and deterministic approach for EL that combines lightweight and easily extensible graph-based algorithms with a new context-based retrieval method.
- MAG features an innovative candidate generation method that relies on various filter methods and search types for a better candidate selection.
- We provide a thorough evaluation of our overall system on 23 datasets using the GERBIL platform (Usbeck et al., 2015). Our results outperform all state-of-the-art

[8] https://github.com/dice-group/AGDISTIS
[9] https://hobbitdata.informatik.uni-leipzig.de/agdistis/

approaches on 6 non-English datasets while achieving state-of-the-art performance on English.

2.2 RDF2PT: GENERATING BRAZILIAN PORTUGUESE TEXTS FROM RDF DATA

FOR MITIGATING THE LACK OF MULTILINGUAL APPROACHES IN TEXT GENERATION 1.1.2, we developed RDF2PT, an approach that verbalizes RDF data to Brazilian Portuguese language. A generic NLG pipeline is composed by three tasks which are *Document Planing*, *Micro Planning* and *Realization*. RDF2PT operates mostly at the level of the first two and to the *Realization* task, RDF2PT uses an adaption of SimpleNLG to Brazilian Portuguese (De Oliveira and Sripada, 2014).

2.2.1 DOCUMENT PLANNING

This initial phase is divided into two sub-tasks. First, *Content determination*, which decides what information a certain NLG system should include in the generated text. Second, *Discourse planning*, which determines the order of the information in paragraphs and its rhetorical relation.

Content determination. RDF2PT assumes the description of a resource to be the set of RDF statements of which this resource is the subject. Hence, given a resource, RDF2PT first performs a SPARQL query to get its most specific class through the predicate `rdf:type`. Afterward, RDF2PT gets all resources that belong to this specific class and ranks their predicates by using Page Rank (Page et al., 1999) over the KB. Once the predicates are ranked, RDF2PT considers only the top seven most popular predicates of the class to describe the input resource.

Discourse planning. In this step, RDF2PT clusters and orders the triples. The subjects are ordered with respect to the number of their occurrences, thus assigning them to those input triples that mention them. RDF2PT processes the input in descending order with respect to the frequency of the variables they contain, starting with the projection variables and only after that, turning to other variables.

2.2.2 MICRO PLANNING

This step is concerned with the planning of a sentence. It comprises three sub-tasks. Firstly, *Sentence aggregation* decides whether information will be presented individually or separately. Second, *Lexicalization* chooses the right words and phrases in natural language for expressing the semantics about the data. Third, *Referring Expression*) is the task responsible for generating syntagms (references) to discourse entities.

Sentence aggregation. This task is divided into two phases, *subject grouping* and *object grouping*. *Subject grouping* collapses the predicates and objects of two triples if their subjects are the same. *Object grouping* collapses the subjects of two triples if the predicates and objects of the triples are the same. The common elements are usually subject noun phrases and verb phrases (verbs together with object noun phrases). To maximize the grouping effects, we additionally collapse common prefixes and suffixes of triples, irrespective of whether they are full subject noun phrases or complete verb phrases.

Lexicalization. This step comprises the main contribution of RDF2PT for verbalizing the triples in Brazilian Portuguese. In contrast to English, Brazilian Portuguese is a morphologically rich language which contains the grammatical gender of words. Grammatical gender plays a key role because it affects the generation of determiners and pronouns. It also influences the inflection of nouns and verbs. For instance, the passive expression of the verb `nascer` (en: “be born") is `nascida` if the subject is feminine or `nascido` if masculine. Thus, the gender of words is essential for comprehending the semantics of a given Portuguese text. Also, Brazilian Portuguese has different possibilities in the expression of subject possessives. Hence, RDF2PT has to deal with the following phenomena while lexicalizing:

- **Grammatical gender** - In Portuguese, the gender varies between masculine and feminine. This variation leads to supplementary challenges when lexicalizing words automatically. For example, a gender may be represented by articles “um" and “o" (masculine) or “uma" and “a" (feminine). However, the gender also affects the inflection of words. For instance, for the word “cantor" (en: “singer"), if the subject is feminine, the word becomes “cantora". However, there are words that do not inflect, e.g., the word “gerente" (en: “manager"). If the subject is a woman, we only refer to it by using the article “a”, i.e., “a gerente". Therefore, there are some challenges to tackle for recognizing the gender and assigning it

correctly. A tricky example to solve automatically is “O Rio de Janeiro é uma cidade" (en: Rio de Janeiro is a city). In this case, the subject is masculine but its complement is feminine. Developing handcrafted rules to handle these phenomena can become a hard task. To deal with this challenge, we use a Part-Of-Speech tagger (TreeTagger in our case) as it retrieves the gender along with the parts of speech. All the obtained genders are attached along with the lexicalizations for supporting the realization step.

- **Classes and resources** - The lexicalization of classes and resources is gathered by using a SPARQL query to get their Portuguese labels through the `rdfs:label` predicate[10]. In case such a label does not exist, we use either the fragment of their URI (the string after the # character) if it exists, or the string after the last occurrence of “/". Finally, this natural language representation is lexicalized as a noun phrase. Afterwards, RDF2PT recognizes the gender. In case the resource is recognized as a person, RDF2PT applies a string similarity measure (0.8 threshold) between the lexicalized word and a list of names provided by LD2NL. This list is divided by masculine and feminine which in turn results in the gender. On the other hand, if the resource is not a person, we use Tree-tagger.

- **Properties** - The lexicalization of properties relies on one of the results of-Ngonga Ngomo et al. (2013), i.e., that most property labels are either nouns or verbs. To determine which lexicalization to use automatically, we rely on the insight that the first and last words of a property label in Portuguese are commonly the key for determining the type of property. We then use the Tree-Tagger to get the part of speech of predicates. Properties whose label begins with a verb are lexicalized as verbs. For example, the predicate `dbo:knownFor`, which Portuguese label is “conhecido por", has the first word identified as an inflection of the verb “conhecer" (en:know). Therefore, we devised a set of rules to capture this behavior.

- **Literals** - In an RDF graph, literals usually consist of a *lexical form* `LF` and a *datatype IRI* `DT`. If the datatype is `rdf:langString`, a non-empty *language*

[10]Note that it could be any property which returns a natural language representation of the given URI, see (Ell et al., 2011).

tag is specified and the literal is denoted as a *language-tagged string*.[11] Accordingly, the lexicalization of strings with language tags is carried out by using simply the lexical form, while omitting the language tag. For example, "`Albert Einstein"@pt` is lexicalized as "Albert Einstein" or "`Alemanha"@pt` ("Germany"@en) is lexicalized as "Alemanha".

REG. In this step, RDF2PT relies on the number of subjects contained by the RDF statements and only uses other expressions to refer to a given subject in case there is more than one mention of it. RDF2PT replaces the subject by possessive or personal pronouns with the corresponding gender depending on the predicates. For instance, given a triple `dbr: Albert_Einstein dbo:birthPlace dbr:Ulm`, the predicate is a noun phrase then the subject is replaced by a possessive form which is "seu" (en:"his"). However, Brazilian Portuguese has two different ways to express possession and this variation exists due to the necessity of handling complex syntaxes in some sentences and also because the gender of pronouns agrees with objects instead of subjects. For example, "A professora proibiu que o aluno utilizasse `seu` dicionário." (eng: "The teacher forbade the student to use `his/her` dictionary"). The possessive pronoun `seu` in this sentence does not indicate explicitly to whom the dictionary belongs, if it belongs to the `professora` (eng:teacher) or `aluno` (eng:student). Thus, we have explicitly to define the possessive pronoun in order to decrease the ambiguity in texts and it is obviously important when generating text from data. If this sentence was translated into English, we would have indicated to whom the dictionary belonged, `her` or `his`. To this end, we handle the ambiguity of possessive pronouns by interspersing the alternative forms, e.g., `dele` (eng:his) or `dela` (eng: her)" that agrees with the subject. However, it is used just in case more than one subject exists in the same description.

2.2.3 LINGUISTIC REALISATION

This last step is responsible for mapping the obtained descriptions of sentences from the aforementioned tasks and verbalizing them syntactically, morphologically and orthographically into a correct natural language text. To this end, we perform this step by relying on a Brazilian adaptation of SimpleNLG (De Oliveira and Sripada, 2014) and Ngonga Ngomo et al. (2013).

[11] In RDF 1.0 literals have been divided into "plain' literals with no type and optional language tags, and typed literals.

2.2.4 EVALUATION

We based our evaluation methodology on Gardent et al. (2017c) and Ferreira et al. (2016). Our main goal was to evaluate how well RDF2PT represents the information obtained from the data. We hence divided our evaluation set into expert and non-expert users. Both sets were made up of native speakers of Brazilian Portuguese. We selected six DBpedia categories like (Gardent et al., 2017c) for selecting the topic of texts. The categories were Astronaut, Scientist, Building, WrittenWork, City, and University.

Experts - We aimed to evaluate the adequacy and fluency of the generated texts from 10 experts. All experts hold at least a master degree in the fields NLP or Semantic Web (SW). In the questionnaire, we used the same two questions as (Gardent et al., 2017c): (1) Adequacy: Does the text contain only and all the information from the data? (2) Fluency: Does the text sound fluent and natural?

Non-experts - We evaluated the clarity and fluency of the generated texts. To this end, we created three types of texts, (1) baseline, (2) RDF2PT and (3) Human. The experiment was performed by 30 participants (10 per list). They were asked to rate each text considering the clarity and fluency based on two questions from Ferreira et al. (2016) on a scale from 1 (Very Bad) to 5 (Very Good). The questions were: (1) Fluency: Does the text present a consistent, logical flow? (2) Clarity: Is the text easy to understand?

In total, we created three versions of 18 texts (one text per resource) selected randomly from the aforementioned DBpedia categories (total: 54 texts). These texts were distributed over three lists, such that each list contained one variant of each text, and there was an equal number of texts from the three types (Baseline, RDF2PT, Human).

2.2.5 RESULTS

Experts Figure 2.4 displays the average fluency and clarity of the texts. The results suggest that RDF2PT is able to capture and represent the information from data adequately. Also, the generated texts are fluent enough to be understood by humans.

Non-experts Figure 2.5 depicts the average fluency and clarity of the texts where their topics are described by *Baseline*, *RDF2PT* and *Human* approaches respectively. This figure clearly shows that *Baseline* texts are rated lower than both the *RDF2PT* and *Human* texts, in fact, *RDF2PT* is superior to *Baseline* and close to *Human*.

We performed a statistical analysis in order to measure the significance of the difference between the types (Baseline, RDF2PT, Human). First, we carried out a Friedman

Version	Text
Baseline	Albert Einstein é cientista, Albert Einstein campo é física, Albert Einstein lugar falecimento Princeton. Albert Einstein ex-instituição é Universidade Zurique, Albert Einstein é conhecido Equivalência massa-energia, Albert Einstein prêmio é Medalha Max Planck, Albert Einstein estudante doutorado é Ernst Gabor Straus.
RDF2PT	Albert Einstein foi um cientista, o campo dele foi a física e ele no Princeton. Além disso, sua ex-instituição foi a Universidade de Zurique, ele é conhecido pela Equivalência massa-energia, o prêmio dele foi a Medalha Max Planck e o estudante de doutorado dele foi o Ernst Gabor Straus.
Humano	Albert Einstein era um cientista, que trabalhava na área de Física. Era conhecido pela fórmula de equivalência entre massa e energia. Formou-se na Universidade de Zurique. Einstein ganhou a medalha Max Planck por seu trabalho. Em Princeton, onde morreu, teve sob sua orientação Ernst Gabor Straus.

Table 2.3: Example of text in the Baseline, RDF2PT approach and Human version.

test (Friedman, 1937) which resulted in a significant difference in the fluency (x^2 = 193.61, ρ <0.0001) and clarity (x^2 = 180.9, ρ <0.0001) for the three kinds of texts. Afterward, we conducted a post-hoc analysis with Wilcoxon signed-rank test corrected for multiple comparisons using the Bonferroni method, resulting in a significance level set at ρ <0.017. Texts of the Baseline are hence significantly less statistically understandable (Z=525 and ρ <0.017.) and fluent (Z=275.5 and ρ <0.017.) than those generated by the RDF2PT approach. However, RDF2PT also generates texts less comprehensible (Z=1617.5 and ρ

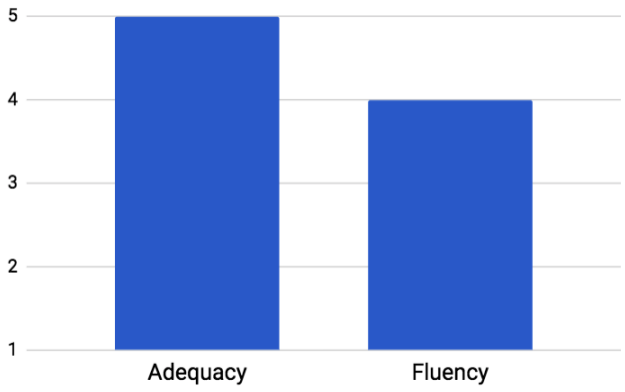

Figure 2.4: RDF2PT results in experts survey

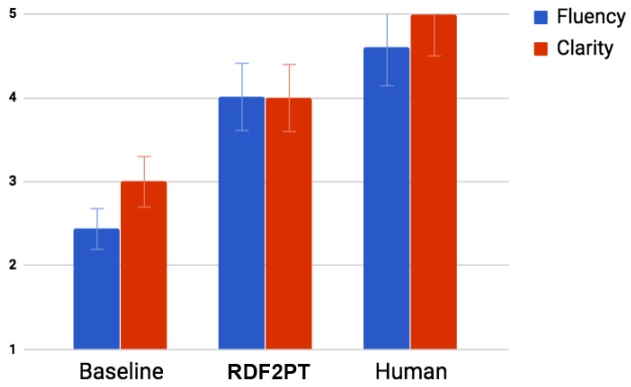

Figure 2.5: Results in non-experts experiment

<0.017.) and fluent (Z=1640.0 and ρ <0.017.) than those generated by humans. Clearly, humans were superior to Baseline in terms of comprehensibility (Z=234.5 and ρ <0.017.) and fluency (Z=264.0.0 and ρ <0.017.), as we expected. Therefore, there is a significant difference among all models, being baseline < model < human.

2.2.6 REPRODUCIBILITY

All experimental data, code, and results are publicly available[12] as well as the extensions to more than one language. [13] In addition, the experiment was run on CrowdFlower and is publicly available.[14]

2.2.7 SUMMARY

The main contributions of this paper can be summarized as follows:

- We present the first RDF-to-Text approach to Brazilian Portuguese.
- RDF2PT is extensible for Spanish and English as well as other languages such as Italian, French and German.
- RDF2PT generates natural language sentences close to the human quality.

[12]https://github.com/dice-group/RDF2PT
[13]https://github.com/diegomoussallem/RDF2NL
[14]https://ilk.uvt.nl/~tcastrof/semPT/evaluation/

2.3 NEURALREG: AN END-TO-END APPROACH TO REFERRING EXPRESSION GENERATION

FOR IMPROVING THE REFERRING FORM OF ENTITIES IN TEXT GENERATION 1.1.2, we created the first approach, named NeuralREG, which relies on deep neural networks for making decisions about form and content in one go without explicit feature extraction from RRDF-KG.

NeuralREG accepts as input entities which are delexicalized to general tags (e.g., ENTITY-1, ENTITY-2) to decrease data sparsity. Based on the delexicalized input, the model generates outputs which may be likened to templates in which references to the discourse entities are not realized (as in "The ground of ENTITY-1 is located in ENTITY-2."). To this end, NeuralREG uses as training data a specific constructed set of 78,901 referring expressions to 1,501 entities in the context of the RDF-KG, derived from a (delexicalized) version of the WebNLG corpus (Gardent et al., 2017a,b).

NeuralREG aims to generate a referring expression $y = \{y_1, y_2, ..., y_T\}$ with T tokens to refer to a target entity token $x^{(wiki)}$ given a discourse pre-context $X^{(pre)} = \{x_1^{(pre)}, x_2^{(pre)}, ..., x_m^{(pre)}\}$ and pos-context $X^{(pos)} = \{x_1^{(pos)}, x_2^{(pos)}, ..., x_l^{(pos)}\}$ with m and l tokens, respectively. The model is implemented as a multi-encoder, attention-decoder network with bidirectional (Schuster and Paliwal, 1997) Long Short-Term Memories (LSTM) (Hochreiter and Schmidhuber, 1997) sharing the same input word-embedding matrix V, we detail it in the next sections.

2.3.1 CONTEXT ENCODERS

Our model starts by encoding the pre- and pos-contexts with two separate bidirectional LSTM encoders (Schuster and Paliwal, 1997; Hochreiter and Schmidhuber, 1997). These modules learn feature representations of the text surrounding the target entity $x^{(wiki)}$, which are used for the referring expression generation. The pre-context $X^{(pre)} = \{x_1^{(pre)}, x_2^{(pre)}, ..., x_m^{(pre)}\}$ is represented by forward and backward hidden-state vectors $(\overrightarrow{h}_1^{(pre)}, \cdots, \overrightarrow{h}_m^{(pre)})$ and $(\overleftarrow{h}_1^{(pre)}, \cdots, \overleftarrow{h}_m^{(pre)})$. The final annotation vector for each encoding timestep t is obtained by the concatenation of the forward and backward representations $h_t^{(pre)} = [\overrightarrow{h}_t^{(pre)}, \overleftarrow{h}_t^{(pre)}]$. The same process is repeated for the pos-context resulting in representations $(\overrightarrow{h}_1^{(pos)}, \cdots, \overrightarrow{h}_l^{(pos)})$ and $(\overleftarrow{h}_1^{(pos)}, \cdots, \overleftarrow{h}_l^{(pos)})$ and annotation vectors $h_t^{(pos)} = [\overrightarrow{h}_t^{(pos)}, \overleftarrow{h}_t^{(pos)}]$. Finally, the encoding of target entity $x^{(wiki)}$ is simply its entry in the shared input word-embedding matrix V_{wiki}.

2.3.2 DECODER

The referring expression generation module is an LSTM decoder implemented in three different versions: `Seq2Seq`, `CAtt` and `HierAtt`. All decoders at each timestep i of the generation process take as input features their previous state s_{i-1}, the target entity-embedding V_{wiki}, the embedding of the previous word of the referring expression $V_{y_{i-1}}$ and finally the summary vector of the pre- and pos-contexts c_i. The difference between the decoder variations is the method to compute c_i.

`Seq2Seq` models the context vector c_i at each timestep i concatenating the pre- and pos-context annotation vectors averaged over time:

$$\hat{h}^{(pre)} = \frac{1}{N}\sum_{i}^{N} h_i^{(pre)} \tag{2.1}$$

$$\hat{h}^{(pos)} = \frac{1}{N}\sum_{i}^{N} h_i^{(pos)} \tag{2.2}$$

$$c_i = [\hat{h}^{(pre)}, \hat{h}^{(pos)}] \tag{2.3}$$

`CAtt` is an LSTM decoder augmented with an attention mechanism (Bahdanau et al., 2014) over the pre- and pos-context encodings, which is used to compute c_i at each timestep. We compute energies $e_{ij}^{(pre)}$ and $e_{ij}^{(pos)}$ between encoder states $h_i^{(pre)}$ and $h_i^{(post)}$ and decoder state s_{i-1}. These scores are normalized through the application of the softmax function to obtain the final attention probability $\alpha_{ij}^{(pre)}$ and $\alpha_{ij}^{(post)}$. Equations 2.4 and 2.5 summarize the process with k ranging over the two encoders ($k \in [pre, pos]$), being the projection matrices $W_a^{(k)}$ and $U_a^{(k)}$ and attention vectors $v_a^{(k)}$ trained parameters.

$$e_{ij}^{(k)} = v_a^{(k)T} \tanh(W_a^{(k)} s_{i-1} + U_a^{(k)} h_j^{(k)}) \tag{2.4}$$

$$\alpha_{ij}^{(k)} = \frac{\exp(e_{ij}^{(k)})}{\sum_{n=1}^{N} \exp(e_{in}^{(k)})} \tag{2.5}$$

In general, the attention probability $\alpha_{ij}^{(k)}$ determines the amount of contribution of the jth token of k-context in the generation of the ith token of the referring expression. In each decoding step i, a final summary-vector for each context $c_i^{(k)}$ is computed by summing the encoder states $h_j^{(k)}$ weighted by the attention probabilities $\alpha_i^{(k)}$:

$$c_i^{(k)} = \sum_{j=1}^{N} \alpha_{ij}^{(k)} h_j^{(k)} \tag{2.6}$$

To combine $c_i^{(pre)}$ and $c_i^{(pos)}$ into a single representation, this model simply concatenates the pre- and pos-context summary vectors $c_i = [c_i^{(pre)}, c_i^{(pos)}]$.

`HierAtt` implements a second attention mechanism inspired by Libovický and Helcl (2017) in order to generate attention weights for the pre- and pos-context summary-vectors $c_i^{(pre)}$ and $c_i^{(pos)}$ instead of concatenating them. Equations 2.7, 2.8 and 2.9 depict the process, being the projection matrices $W_b^{(k)}$ and $U_b^{(k)}$ as well as attention vectors $v_b^{(k)}$ trained parameters ($k \in [pre, pos]$).

$$e_i^{(k)} = v_b^{(k)T} \tanh(W_b^{(k)} s_{i-1} + U_b^{(k)} c_i^{(k)}) \tag{2.7}$$

$$\beta_i^{(k)} = \frac{\exp(e_i^{(k)})}{\sum_n \exp(e_i^{(n)})} \tag{2.8}$$

$$c_i = \sum_k \beta_i^{(k)} U_b^{(k)} c_i^{(k)} \tag{2.9}$$

`Decoding` Given the summary-vector c_i, the embedding of the previous referring expression token $V_{y_{i-1}}$, the previous decoder state s_{i-1} and the entity-embedding V_{wiki}, the decoders predict their next state which is used later to compute a probability distribution over the tokens in the output vocabulary for the next timestep as Equations 2.10 and 2.11 show.

$$s_i = \Phi_{\text{dec}}(s_{i-1}, [c_i, V_{y_{i-1}}, V_{wiki}]) \tag{2.10}$$

$$p(y_i | y_{<i}, X^{(pre)}, x^{(wiki)}, X^{(pos)}) = \text{softmax}(W_c s_i + b) \tag{2.11}$$

In Equation 2.10, s_0 and c_0 are zero-initialized vectors. In order to find the referring expression y that maximizes the likelihood in Equation 2.11, we apply a beam search with length normalization with $\alpha = 0.6$ (Wu et al., 2016):

$$lp(y) = \frac{(5 + |y|)^\alpha}{(5 + 1)^\alpha} \tag{2.12}$$

The decoder is trained to minimize the negative log likelihood of the next token in the target referring expression:

$$J(\theta) = -\sum_i \log \text{p}(y_i | y_{<i}, X^{(pre)}, x^{(wiki)}, X^{(pos)}) \tag{2.13}$$

2.3.3 AUTOMATIC EVALUATION

Data - We evaluated our models on the training, development and test referring expression sets of the delexicalized WebNLG.

Metrics - We compared the referring expressions produced by the evaluated models with the gold-standards ones using accuracy and String Edit Distance (Levenshtein, 1966). Since pronouns are highlighted as the most likely referential form to be used when a referent is salient in the discourse, as argued in the introduction, we also computed pronoun accuracy, precision, recall and F1-score in order to evaluate the performance of the models for capturing discourse salience. Finally, we lexicalized the original templates with the referring expressions produced by the models and compared them with the original texts in the corpus using accuracy and BLEU score (Papineni et al., 2002a) as a measure of fluency. Since our model does not handle referring expressions for constants (dates and numbers), we just copied their source version into the template.

Post-hoc McNemar's and Wilcoxon signed ranked tests adjusted by the Bonferroni method were used to test the statistical significance of the models in terms of accuracy and string edit distance, respectively. To test the statistical significance of the BLEU scores of the models, we used a bootstrap resampling together with an approximate randomization method (Clark et al., 2011)[15].

Settings - NeuralREG was implemented using Dynet (Neubig et al., 2017). Source and target word embeddings were 300D each and trained jointly with the model, whereas hidden units were 512D for each direction, totaling 1024D in the bidirection layers. All non-recurrent matrices were initialized following the method of Glorot and Bengio (2010). Models were trained using stochastic gradient descent with Adadelta (Zeiler, 2012) and mini-batches of size 40. We ran each model for 60 epochs, applying early stopping for model selection based on accuracy of the development set with patience of 20 epochs. For each decoding version (`Seq2Seq`, `CAtt` and `HierAtt`), we searched for the best combination of drop-out probability of 0.2 or 0.3 in both the encoding and decoding layers, using beam search with a size of 1 or 5 with predictions up to 30 tokens or until 2 ending tokens were predicted (*EOS*). The results described in the next section were obtained on the test set by the NeuralREG version with the highest accuracy on the development set over the epochs.

[15]https://github.com/jhclark/multeval

2.3.4 HUMAN EVALUATION

Complementary to the automatic evaluation, we performed an evaluation with human judges, comparing the quality judgments of the original texts to the versions generated by our various models.

Material - We quasi-randomly selected 24 instances from the delexicalized version of the WebNLG corpus related to the test part of the referring expression collection. For each of the selected instances, we took into account its source triple set and its 6 target texts: one original (randomly chosen) and its versions with the referring expressions generated by each of the five models introduced in this study (two baselines, three neural models). Instances were chosen following two criteria: the number of triples in the source set (ranging from 2 to 7) and the differences between the target texts.

For each size group, we randomly selected four instances (of varying degrees of variation between the generated texts) giving rise to 144 trials (= 6 triple set sizes $*$ 4 instances $*$ 6 text versions), each consisting of a set of triples and a target text describing it with the lexicalized referring expressions highlighted in yellow.

Method - The experiment had a latin-square design, distributing the 144 trials over 6 different lists such that each participant rated 24 trials, one for each of the 24 corpus instances, making sure that participants saw equal numbers of triple set sizes and generated versions. Once introduced to a trial, the participants were asked to rate the fluency ("does the text flow in a natural, easy to read manner?"), grammaticality ("is the text grammatical (no spelling or grammatical errors)?") and clarity ("does the text clearly express the data?") of each target text on a 7-Likert scale, focussing on the highlighted referring expressions. The experiment is available on the website of the author[16].

Participants - We recruited 60 participants, 10 per list, via Mechanical Turk. Their average age was 36 years and 27 of them were females. The majority declared themselves native speakers of English (44), while 14 and 2 self-reported as fluent or having a basic proficiency, respectively.

2.3.5 RESULTS.

Automatic evaluation - Table 2.4 summarizes the results for all models on all metrics on the test set and Table 2.5 depicts a text example lexicalized by each model. The first thing to note in the results of the first table is that the baselines in the top two rows performed

[16]https://ilk.uvt.nl/~tcastrof/acl2018/evaluation/

	All References		Pronouns				Text	
	Acc.	SED	Acc.	Prec.	Rec.	F-Score	Acc.	BLEU
OnlyNames	0.53^{D}	4.05^{D}	-	-	-	-	0.15^{D}	69.03^{D}
Ferreira	0.61^{C}	3.18^{C}	0.43^{B}	0.57	0.54	0.55	0.19^{C}	72.78^{C}
NeuralREG+`Seq2Seq`	$0.74^{A,B}$	$2.32^{A,B}$	0.75^{A}	0.77	0.78	0.78	0.28^{B}	$79.27^{A,B}$
NeuralREG+`CAtt`	0.74^{A}	2.25^{A}	0.75^{A}	0.73	0.78	0.75	0.30^{A}	79.39^{A}
NeuralREG+`HierAtt`	0.73^{B}	2.36^{B}	0.73^{A}	0.74	0.77	0.75	$0.28^{A,B}$	79.01^{B}

Table 2.4: (1) Accuracy (Acc.) and String Edit Distance (SED) results in the prediction of all referring expressions; (2) Accuracy (Acc.), Precision (Prec.), Recall (Rec.) and F-Score results in the prediction of pronominal forms; and (3) Accuracy (Acc.) and BLEU score results of the texts with the generated referring expressions. Rankings were determined by statistical significance.

quite strong on this task, generating more than half of the referring expressions exactly as in the gold-standard. The method based on Castro Ferreira et al. (2016) performed statistically better than *OnlyNames* on all metrics due to its capability, albeit to a limited extent, to predict pronominal references (which *OnlyNames* obviously cannot).

We reported results on the test set for NeuralREG+`Seq2Seq` and NeuralREG+`CAtt` using dropout probability 0.3 and beam size 5, and NeuralREG+`HierAtt` with dropout probability of 0.3 and beam size of 1 selected based on the highest accuracy on the development set. Importantly, the three NeuralREG variant models statistically outperformed the two baseline systems. They achieved BLEU scores, text and referential accuracies as well as string edit distances in the range of 79.01-79.39, 28%-30%, 73%-74% and 2.25-2.36, respectively. This means that NeuralREG predicted 3 out of 4 references completely correct, whereas the incorrect ones needed an average of 2 post-edition operations in character level to be equal to the gold-standard. When considering the texts lexicalized with the referring expressions produced by NeuralREG, at least 28% of them are similar to the original texts. Especially noteworthy was the score on pronoun accuracy, indicating that the model was well capable of predicting when to generate a pronominal reference in our dataset.

The results for the different decoding methods for NeuralREG were similar, with the NeuralREG+`CAtt` performing slightly better in terms of the BLEU score, text accuracy and String Edit Distance. The more complex NeuralREG+`HierAtt` yielded the lowest results, even though the differences with the other two models were small and not even statistically significant in many of the cases.

Model	Text
OnlyNames	**alan shepard** was born in **new hampshire** on **1923-11-18** . before **alan shepard** death in **california alan shepard** had been awarded **distinguished service medal (united states navy)** an award higher than **department of commerce gold medal** .
Ferreira	**alan shepard** was born in **new hampshire** on **1923-11-18** . before **alan shepard** death in **california him** had been awarded **distinguished service medal** an award higher than **department of commerce gold medal** .
`Seq2Seq`	**alan shepard** was born in **new hampshire** on **1923-11-18** . before **his** death in **california him** had been awarded **the distinguished service medal by the united states navy** an award higher than **the department of commerce gold medal** .
`CAtt`	**alan shepard** was born in **new hampshire** on **1923-11-18** . before **his** death in **california** **he** had been awarded **the distinguished service medal by the us navy** an award higher than **the department of commerce gold medal** .
`HierAtt`	**alan shephard** was born in **new hampshire** on **1923-11-18** . before **his** death in **california he** had been awarded **the distinguished service medal** an award higher than **the department of commerce gold medal** .
Original	**alan shepard** was born in **new hampshire** on **18 november 1923** . before **his** death in **california he** had been awarded **the distinguished service medal by the us navy** an award higher than **the department of commerce gold medal** .

Table 2.5: Example of text with references lexicalized by each model.

Human evaluation - Table 2.6 summarizes the results and reveals a clear pattern: all three neural models scored higher than the baselines on all metrics, with especially NeuralREG+`CAtt` approaching the ratings for the original sentences, although differences between the neural models were small. Concerning the size of the triple sets, we did not find any clear pattern. To test the statistical significance of the pairwise comparisons, we used the Wilcoxon signed-rank test corrected for multiple comparisons using the Bonferroni method. In contrast to the automatic evaluation, the results of both baselines were not statistically significant for the three metrics. In comparison with the neural models, NeuralREG+`CAtt` significantly outperformed the baselines in terms of fluency, whereas the other comparisons between baselines and neural models were not statistically significant. The results for the three different decoding methods of NeuralREG did not reveal a significant difference as well. Finally, the original texts were rated significantly higher than both baselines in terms of the three metrics, also than NeuralREG+`Seq2Seq` and NeuralREG+`HierAtt` in terms of fluency, and higher than NeuralREG+`Seq2Seq` in terms of clarity.

	Fluency	Grammar	Clarity
OnlyNames	4.74^{C}	4.68^{B}	4.90^{B}
Ferreira	4.74^{C}	4.58^{B}	4.93^{B}
NeuralREG+`Seq2Seq`	$4.95^{B,C}$	$4.82^{A,B}$	4.97^{B}
NeuralREG+`CAtt`	$5.23^{A,B}$	$4.95^{A,B}$	$5.26^{A,B}$
NeuralREG+`HierAtt`	$5.07^{B,C}$	$4.90^{A,B}$	$5.13^{A,B}$
Original	5.41^{A}	5.17^{A}	5.42^{A}

Table 2.6: Fluency, Grammaticality and Clarity results obtained in the human evaluation. Rankings were determined by statistical significance.

2.3.6 REPRODUCIBILITY

All experimental data, code, and models are publicly available.[17].

2.3.7 SUMMARY

The main contributions of this paper can be summarized as follows:

- We present the first full REG model based on NN from RDF-KG.
- NeuralREG achieves fluency close to the human quality while choosing the referential form of entities in text.
- NeuralREG achieves high accuracy in predicting the form of pronouns for entities.

2.4 KG-NMT: UTILIZING KNOWLEDGE GRAPHS FOR NEURAL MACHINE - TRANSLATION AUGMENTATION

FOR ADDRESSING THE ENTITY TRANSLATION PROBLEM IN TEXT 1.1.3, we designed KG-NMT, an NMT model which is augmented with KG. KG-NMT is based on the observation that more than 150 billion facts referring to more than 3 billion entities are available in the form of KG on the Linked Open Data (LOD) Cloud.[18] Hence, the intuition behind our methodology is as follows: *Given that KGs describe real-world entities, we can use a KG along with EL to optimize values of entities' vector in the*

[17]https://github.com/ThiagoCF05/NeuralREG
[18]http://lod-cloud.net/

embedding space and consequently to achieve a better translation quality of entities in text. Figure 2.6 depicts the general idea of our methodology.

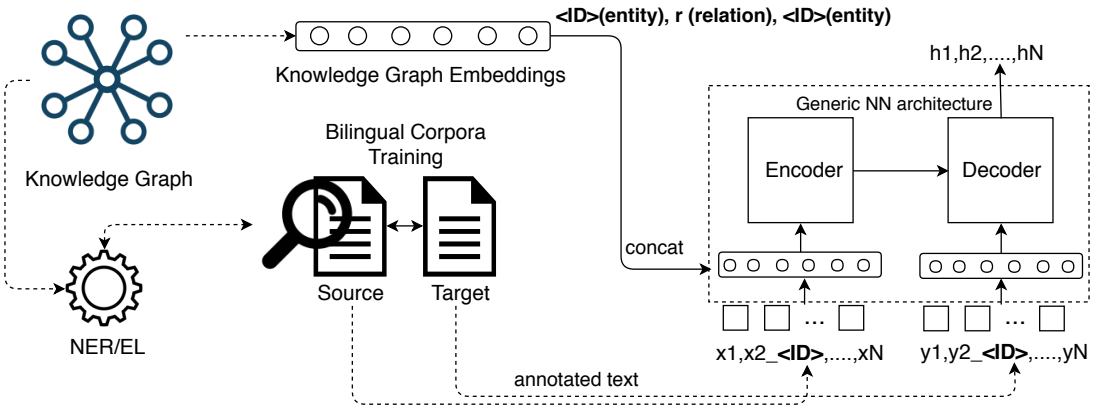

Figure 2.6: Overview of the KG-NMT methodology.

We devised two strategies to instantiate our methodology. In the first strategy, we link the NEs in the source and target texts to a reference KB, i.e., DBpedia, using MAG (Moussallem et al., 2017), a multilingual EL system. We then incorporate the URIs of entities along with the tokens akin to Li et al. (2018) with the NE-tags. For example, the word *cancer* can be annotated with `cancer|dbr_Cancer`,[19] and its translation represented in the German part of the DBpedia KB (`dbr_Krebs_(Medizin)`). After incorporating the URIs, we embed the reference KB, using the *fastText* KGE algorithm (Joulin et al., 2017). Once the KGEs are created, we concatenate their vectors to the internal vectors of NMT embeddings. The concatenation is possible as the annotations are present in the texts and consequently in the vocabulary. We chose to concatenate the vectors instead of leveraging their values because the concatenation preserves the values of KGEs while leveraging loses its original values. In case a given entity in the text does not have a vector in the KGEs, the concatenation inserts an empty vector, while the performance of NMT remains unaffected. For example, suppose the entity "USA" appears in the parallel training data, i.e., a 500 dimensional vectors space is learned. Likewise, it also appears in the KG, thus a vector space of the same dimension is learned. After concatenation, the *USA* embedding becomes 2x500 dimensions, whereby entities, which are not present in the KGEs are concatenated with an empty 500 dimensional vector.

[19]http://dbpedia.org/resource/Cancer

Although incorporating EL as a feature into NMT is interesting by itself, the annotation of entities in the training set and the post-editing task can be resource-intensive. Additionally, one limitation of structure-based KGEs is that it can only work with word-based models since the application of any segmentation model, such as BPE, on entities and relations may force the algorithm to assign wrong vectors to the entities. BPE is a form of data compression that iteratively replaces the most frequent pair of bytes in a sequence with a single, unused byte. For example, the entities `dbr:Leipzig` and `dbr:Leibniz`[20] can be similar when considering sub-word units (characters), however, the first is a location while the second is a person. Both entities can be connected via the entity `dbr:University_of_Leipzig` but in this case, we are analyzing the sub-word units and not their graph connections. Thus, they should not be regarded as similar in the perspective of entities.

To overcome this limitation, we devise our second strategy, which uses only semantically-enriched KGEs and skips the EL part. Here, we enrich the structure-based KGEs with surface forms of the entities found in the DBpedia KB, thus decreasing the annotation effort and allowing the use of segmentation-models, i.e, sub-word information, on the surface forms differently from the first strategy, which considered only entities. Our hypothesis lies in the unsupervised learning capability of NNs that can predict the annotation and alignment of the entities by itself. To generate the semantically-enriched KGEs, we rely on multinomial logistic regression (Böhning, 1992) as a classifier in a supervised training implemented in *fastText* which assigns labels to the vector representations. The classification task creates inverse relations among the resources in order to map the relation of the subjects and objects in the triples and also it assigns a label, in this case, the entity's URI, to the surface forms of entities and include them in the same vector space. The goal of the task is to predict the URIs of entities by their surface form. For example, we add to the triple, `<USA, type, Country>` the following information, `<USA, surfaceForm, United States of America>`. Thus, the training data looks `__label__dbr:USA United States America`.[21] The classifier creates an additional (hidden) relational vector between every entity (labels) and their associated words. Thus, the model when asked for the label of "United States of America" returns `dbr:USA`. However, we do not rely on the complete model rather we use the KGE generated output with the surface forms attached to it. The semantically-enriched-KGEs

[20] Subword segmentation (BPE) - Leipzig: Le■ ip■ zig ; Leibniz: Le■ ib■ n■ iz

[21] More than one surface forms can be assigned to the entities.

training jointly embeds entities and words into the same vector space, thus generating a vector for every word, which composes the entity's label. Therefore, while learning the KGE along with the NMT vocabulary, the NN can retrieve from the lookup table the surface forms of entities and use their labels URIs to align both source and target entities. By enriching the KGEs with surface forms, it allows using these vectors to initialize the embedding layer's weights of the NMT models. This initialization is similar to the one used with pre-trained monolingual embeddings in NMT (Neishi et al., 2017). But, instead of containing various words, the semantically-enriched KGEs has only the surface forms of the entities, which are also present in the NMT vocabulary. The default initialization of the embeddings layer is a function that assigns random values to the weight matrix, whereas, in our second strategy, the values from KGEs matrix are used to assign constant values to the matrix using a default concat function. Moreover, the employment of semantically-enriched KGEs prevents some errors from the alignment of entities from source and target texts. For example, the entity linker runs in two distinct instances for each language, thus in case a certain entity is annotated only on the source or on the target language side, the NMT approach is affected as the translation task requires aligned bilingual parallel texts for training.

2.4.1 EVALUATION

Different NN architectures are hard to compare as they are susceptible to hyper-parameters. Therefore, we follow the idea of using a minimal reasonable configuration set to the NMT in order to fairly analyze the contributions of the used KG.

NMT Framework - For our overall experiments, we used a bi-directional Recurrent Neural Network (RNN)-LSTM 2-layer encoder-decoder model with attention mechanism (Bahdanau et al., 2014). The training uses a batch size of 32 and the stochastic gradient descent with an initial learning rate of 0.0002. We set a source and target word embeddings' size of 500, and hidden layers to size 500, dropout = 0.3 (naive). We used a maximum sentence length of 80, a vocabulary of 50,000 words for the word based models and a beam size of 5. All experiments were performed with OpenNMT (Klein et al., 2017). In addition, we used a copy mechanism for investigating the OOV words issue. Moreover, we encoded words using BPE with 32,000 merge operations to achieve an open vocabulary. Therefore, we created three kinds of baseline models (word-based, copyM, BPE32) using all the options mentioned above for evaluating the quality of the

translation models. For training the NMT models, we attempted to be as generic as possible. Our training set consists of a merge of the initial one-third of JRC-Acquis 3.0 (Steinberger et al., 2006), Europarl (Koehn, 2005), and OpenSubtitles2013 (Tiedemann, 2012), obtaining a parallel training corpus of two million sentences, containing around 38M running words. We performed our experiments on the English-German language pair as it is one of the most evaluated language pairs in the evaluation campaigns and translating into German is challenging in itself, due to its complex morphology and compounding.

NMT Augmentation - For augmenting the three baseline models with our two KG-based strategies, we annotated the parallel bilingual corpora with MAG (first strategy), a multilingual EL system (Moussallem et al., 2017), which is language and KB agnostic. Afterwards, we trained the KGEs, with a vector dimension size of 500 and a window size of 50 using hierarchical softmax. For semantically-enriched KGEs, we added the surface forms whereby we used the same BPE models on it. For the sake of comparison, we dubbed the KG-NMT approach that relies on EL and structured-based KGEs (first strategy) as *KG-NMT (EL+KGE)* and the version with semantic information (second strategy) as *KG-NMT (SemKGE)*. For overcoming both limitations of *fastText*, which are having a clean KB and information of local graph connectivity, we relied on specific sub-sets of the English and Germany DBpedia KG which contain transitive and CBD resources along with their surface forms.[22] The English KB contains 4.2 million entities, 661 relations, and 2.1 million surface forms, where the German version has 1 million entities, 249 relations, and 0.5 million surface forms.

Evaluation Metrics - We used three automatic MT standard metrics, BLEU (Papineni et al., 2002b), METEOR (Banerjee and Lavie, 2005) and CHRF3 (Popović, 2017) to ensure a consistent and clear evaluation on the common evaluation datasets of the WMT evaluation shared tasks, named *newstest*, between 2015 and 2018 as well as on the domain-specific datasets. Moreover, we carried out a manual analysis of outputs for assuring the contribution from KGs, DBpedia, and we investigated the use of KG in other settings as follows:

Monolingual Embeddings vs. KGEs - Here, we aim to compare the performance of an NMT using pre-trained monolingual embeddings with the semantically-enriched KGEs as both can be used to initialize the internal vectors' values of an NMT model. Our focus is to analyze if the KGEs with fewer vectors can perform better than the

[22]The files are mappingbased_objects, labels, and interlinking_languages

monolingual embeddings for addressing the translation of entities and terminologies. Commonly, pre-trained monolingual word embeddings are used when the bilingual training data of a given language pair is scarce. These monolingual embeddings can be used to maximize the vector values of both, source or target languages as they are usually trained on a large monolingual corpus. Thus, we used a pre-trained monolingual embeddings model, which has 9.2 billion words for English and another with 1.3 billion words for German from (Grave et al., 2018). We dubbed as *biRNN+MonoE*, the NMT model which is maximized with the pre-trained monolingual embeddings.

Continuous Training on Domain-Specific Parallel Datasets - Our goal is to inspect the capability of improving the domain-specific translations using KGs since they document domain-specific information. To this end, we relied on the continued training technique (Luong and Manning, 2015) to adapt a generic NMT system to the financial, medical and Information Technology (IT) domains. For the financial domain, we used the International Financial Reporting Standards (IFRS) ontology and divided the documented labels into training (for continued training), development and test set, containing 1,000 labels each. Similarly, we used the International Classification of Diseases (ICD)-10 ontology, with 1,000 labels in the medical domain for the continued training, development and test set. Finally, for the IT domain, we used the IT-WMT16[23] sets. The continued training set contains 50,121, the development 2,000 and test 1,000 sentences. We followed the same methodology to insert the DBpedia KG in the domain adapted models. The adapted models are named *KG-NMT(EL+KGE)_adapt* and *KG-NMT(SemKGE)_adapt* respectively.

2.4.2 RESULTS

In this section, we report the results of our experiments and perform a manual analysis of each experimental setting.

Overall Results - Table 2.7 shows the results for *KG-NMT* models in comparison to the baselines on the *newstest* dataset between 2015 and 2018. Using KGEs leads to a clear improvement over the baseline as it significantly improved the translation quality in terms of BLEU (+3), METEOR (+4) and CHRF3 (+3) metrics. *KG-NMT (SemKGE)* outperformed *KG-NMT (EL+KGE)* by around +1.3 in BLEU and chrF3, while we observed a +2 point improvement for METEOR. This difference between

[23]http://www.statmt.org/wmt16/it-translation-task.html

Table 2.7: Results in BLEU, METEOR, chrF3 on WMT newstest datasets.

Models		newstest2015			newstest2016			newstest2017			newstest2018		
		BLEU	*METEOR*	*chrF3*	*BLEU*	*METEOR*	*chrF3*	*BLEU*	*METEOR*	*chrF3*	*BLEU*	*METEOR*	*chrF3*
Word-based	biRNN-lstm baseline	16.77	35.20	41.11	18.55	36.62	42.54	15.10	33.75	39.52	20.53	39.02	43.92
	KG-NMT(EL+KGE)	19.86	38.25	42.92	22.38	40.40	45.18	18.04	36.94	41.55	24.87	43.49	46.88
	KG-NMT(SemKGE)	**21.49**	**40.19**	**44.72**	**24.01**	**42.47**	**46.84**	**19.66**	**38.89**	**43.11**	**27.02**	**45.77**	**48.70**
CopyM	biRNN-lstm baseline	19.63	39.20	46.38	21.37	40.90	47.85	17.88	37.89	44.85	24.22	43.96	50.15
	KG-NMT(EL+KGE)	22.46	41.67	48.28	25.05	44.23	50.66	20.77	40.58	47.04	28.44	47.86	53.25
	KG-NMT(SemKGE)	**24.08**	**43.43**	**49.72**	**26.70**	**46.08**	**52.05**	**22.30**	**42.37**	**48.36**	**30.55**	**49.92**	**54.71**
BPE	biRNN-lstm baseline	15.89	36.51	45.97	21.95	42.88	52.68	16.80	39.12	49.35	23.85	45.85	54.98
	KG-NMT(EL+KGE)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	KG-NMT(SemKGE)	**21.74**	**41.41**	**50.04**	**24.86**	**44.32**	**53.59**	**20.45**	**40.62**	**49.45**	**28.02**	**47.51**	**55.16**

the contribution of KGE types is directly related to the EL performance, which did not manage to annotate all kind of entities present in the text. Consequently, the RNN was not able to learn the translations of entities from the DBpedia KG. A different EL that is able to disambiguate more types of entities can improve the results of *KG-NMT (EL+KGE)*, but still its training time is considerable longer than *KG-NMT (SemKGE)*. The augmented model on BPE, *KG-NMT (SemKGE)*, presented consistent improvements showing that the model was capable of learning the segmentation applied on surface forms when translating to morphologically complex languages, such as German. We observed that the copy mechanism can be beneficial for named entities, which were not found in KG. These entities were copied from the source language and added consequently as a translation to the target language. For example, *Chad Johnston* was copied from the source into the target languages as a translation, since this name was not found in the KB.

A detailed study of our results showed that the number of OOV words decreased considerably with the augmentation through KGEs. Table 2.8 shows the number of OOV words generated by the models across all WMT *newstest* datasets. The statistics cannot ensure that every OOV word that became a known word was essentially an entity presented in DBpedia KG. Thus, we chose the *newstest2015* for a manual analysis.[24] We observed that many OOV words, that became known were in fact entities contained

[24] our full analysis of the entities' translation can be found in https://git.io/KG-NMT-experiments-entities

in the KG. As an example (newstest2015 line 1265), the acronym *UK* was not translated by the *biRNN-lstm* baseline even when the copy mechanism (*UK*) or BPE (*Britische*) was used. However, it was correctly translated into German as *Großbritannien* by both KGEs augmented models. Similarly, the entity *Coastguard* (line 1540) was not translated correctly by the baseline models, whereby both KGEs models were able to translate it correctly into *Küstenwache*. Moreover, *KG-NMT (EL+KGE)* was able to translate the word *teacher* (line 438) correctly into *Lehrer* using the knowledge acquired from KG.[25] This human evaluation confirms our hypothesis by showing that the KG-augmented RNN models were able to correctly learn the translation of entities through the relations found in KGEs and that *KG-NMT (SemKGE)* improved generally the translation quality in comparison to other NMT models, which concludes that a supervised annotation of entities with EL is not entirely necessary.

Table 2.8: Number of OOV words across a baseline NMT model and KG-NMT models on the *newstest* dataset.

Models	**2015**	**2016**	**2017**	**2018**
biRNN-lstm baseline	6,004	9,559	9,707	9,383
KG-NMT(EL+KGE)	4,427	6,524	6,603	6,914
KG-NMT(SemKGE)	4,067	5,990	6,130	6,236

Table 2.9: Comparison between pre-trained monolingual embeddings and KGEs.

Models		**newtest2015**			**newtest2016**			**newtest2017**			**newtest2018**		
		BLEU	*METEOR*	*chrF3*	*BLEU*	*METEOR*	*chrF3*	*BLEU*	*METEOR*	*chrF3*	*BLEU*	*METEOR*	*chrF3*
Word-based	biRNN-lstm+MonoE	**21.59**	**40.54**	**45.37**	**24.12**	**42.82**	**47.37**	**20.05**	**39.42**	**43.90**	**27.15**	**46.13**	**49.35**
	KG-NMT(SemKGE)	21.49	40.19	44.72	24.01	42.47	46.84	19.66	38.89	43.11	27.02	45.77	48.70
CopyM	biRNN-lstm+MonoE	**24.21**	**43.81**	**50.32**	**26.97**	**46.52**	**52.61**	**22.61**	**42.87**	**49.01**	**30.77**	**50.39**	**55.41**
	KG-NMT(SemKGE)	24.08	43.43	49.72	26.70	46.08	52.05	22.30	42.37	48.36	30.55	49.92	54.71
BPE	biRNN-lstm+MonoE	19.65	39.24	47.58	**25.13**	**44.66**	**53.54**	**20.93**	**41.41**	**50.33**	**28.42**	**48.00**	**55.98**
	KG-NMT(SemKGE)	**21.74**	**41.41**	**50.04**	24.86	44.32	52.59	20.45	40.62	49.45	28.02	47.51	55.16

Monolingual Embeddings vs KGEs - Table 2.9 reports no significant difference between monolingual embeddings and KGEs in terms of BLEU, METEOR and CHRF3.

[25]http://dbpedia.org/resource/Teacher

Table 2.10: Results of models in BLEU, METEOR, chrF3 on domain-specific testsets.

Models		**ICD-10**			**IFRS**			**IT**		
		BLEU	*METEOR*	*chrF3*	*BLEU*	*METEOR*	*chrF3*	*BLEU*	*METEOR*	*chrF3*
Word-based models	biRNN-lstm baseline_adapt	15.31	23.27	29.63	**52.59**	**60.59**	**62.04**	11.57	28.04	30.50
	KG-NMT(EL+KGE)_adapt	**21.08**	**31.07**	36.93	52.38	60.55	61.86	21.78	40.29	42.75
	KG-NMT(SemKGE)_adapt	20.79	30.70	**37.00**	51.58	59.18	60.05	**23.41**	**41.71**	**44.42**
CopyM models	biRNN-lstm baseline_adapt	16.59	26.49	39.12	**52.91**	**61.68**	**64.34**	13.87	31.61	36.10
	KG-NMT(EL+KGE)_adapt	**22.59**	**34.54**	**46.89**	52.72	61.97	64.65	25.31	44.24	48.96
	KG-NMT(SemKGE)_adapt	22.24	34.10	46.74	51.91	60.33	62.51	**26.84**	**45.81**	**50.38**
BPE Models	biRNN-lstm baseline_adapt	**41.98**	**50.81**	**66.87**	**66.74**	**74.83**	**84.52**	27.83	47.01	**55.91**
	KG-NMT(EL+KGE)_adapt	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	KG-NMT(SemKGE)_adapt	41.44	50.54	66.12	66.21	74.45	84.07	**28.04**	**47.30**	55.68

This finding is interesting since the monolingual embeddings contain billions of words, compared to the DBpedia KG with 4.2 million entities. Taking a deeper look, our manual analysis showed that the OOV words addressed by the monolingual embeddings were not in fact entities, but common words and the entities remained unknown. As an example, the *RNN+MonoE* model translated incorrectly the entity *Principal* into *Wichtigste*, while the *KG-NMT (SemKGE)* used the knowledge documented in the KGs.[26] Moreover, *RNN+MonoE* was unable to translate the entities *UK* and *Coastguard* while the *KG-NMT (SemKGE)* generated the right translations. Therefore, KGEs leverage the world knowledge better than pre-trained monolingual word embeddings for translating entities and we envisage that a combination of both is promising and may lead to further translation improvements.

Continuous Training on Domain-Specific Parallel Datasets - Table 2.10 shows that the knowledge documented in KGs is able to improve significantly the word-based models +5 BLEU, METEOR and chrF3 on the ICD-10 ontology and IT domain. However, no improvement is seen in the IFRS ontology (Financial domain) and all BPE models. Investigating the data and results manually, we perceived that although the terminological expressions infrequently appear in the DBpedia KG, e.g, 14,051 entities in the medical domain in comparison to 4.2 million in the whole graph, its application in the word-based models improved fairly the translations. The same applies to the IT domain, where

[26]http://dbpedia.org/resource/Principal_(school) to translate the entity correctly

the evaluation metric calculates a BLEU score of 11.57 for the baseline system and 23.41 for the semantically-enriched KGEs. However, the lack of improvement in the IFRS ontology was caused by the in-domain training data used in the continued training (domain adaptation). For example, the IFRS data used for continuous training already contained terminological expressions and therefore the adapted models ignored the values from the KGEs. Differently, in the in-domain training of ICD-10 ontology and IT data, the terminological expressions do not appear. Therefore, the IFRS adapted models ignored the values from the KGEs. For this reason, no improvement is seen in the word-based models for the financial domain. Regarding the BPE models, the explanation lies in the capability of NNs for estimating the translation of rarely seen terminological expressions when BPE is applied. Basically, the operations applied by BPE were not common due to very specific NE, for example, names of diseases, for which the NMT system was incapable of learning the translations from the KGEs. In summary, KGEs contribute to the translation of very domain-specific data. However, a further investigation of BPE models in combination with KG-NMT methods is required, as they did not show consistent improvements across the targeted domains.

2.4.3 REPRODUCIBILITY

All experimental data, code, and model are publicly available.[27]

2.4.4 SUMMARY

The main contributions of this paper can be summarized as follows:

- We present the first KG-augmented NMT model, named KG-NMT.
- KG-NMT proposes two strategies for incorporating KGs into NMT models with consistent improvements over baseline.
- KG-NMT shows that KGE leverages better real world knowledge, entities, in comparison two large pre-trained word embeddings trained on large corpora.
- KG-NMT is also capable of translating domain-specific dataset and ontologies.

[27] https://github.com/dice-group/KG-NMT

2.5 THOTH: NEURAL TRANSLATION AND ENRICHMENT OF KNOWLEDGE GRAPHS

FOR ALLEVIATING THE LACK OF MULTILINGUAL KG 1.1.4, we proposed THOTH, an approach for translating and enriching knowledge graphs by relying on neural models. The underlying idea behind our approach, THOTH, is based on the formal description of a translation problem as follows: *Given that KGs are composed of facts extracted from text, we can consider the facts (i.e., triples) as sentences, where URIs are tokens and train a NMT model to translate the facts from one language into another.* The enrichment process implemented by THOTH consists of two phases: the training phase and the translation phase. The data gathering and preprocessing steps occur in the training phase, while the enrichment per se is carried out during the translation phase and consists of two steps: 1) translation and 2) enrichment. All steps carried out in THOTH are language-agnostic, which allow the use of other language-based KGs. An overview can be found in Figure 2.7.

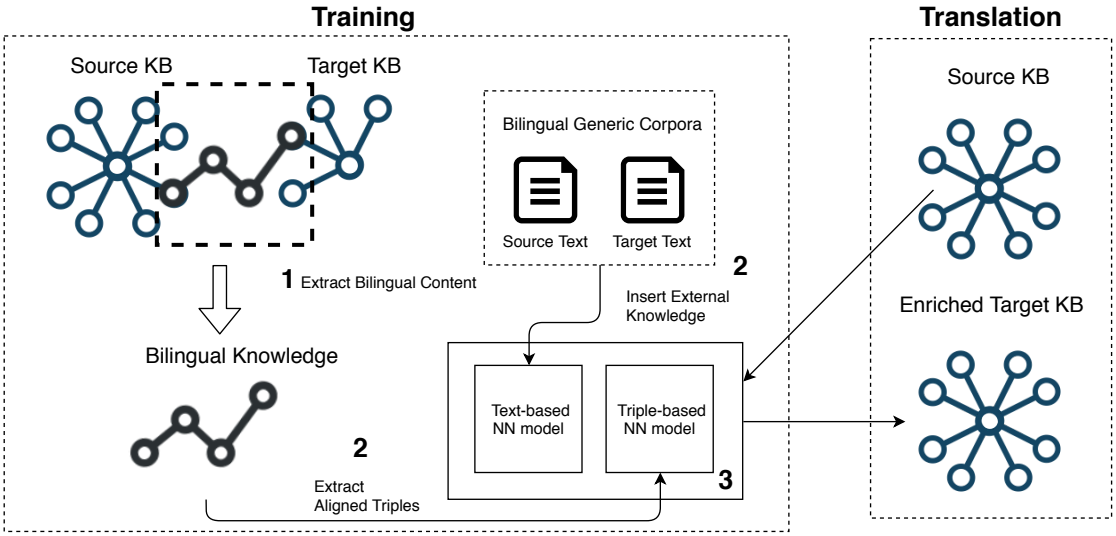

Figure 2.7: Overview of THOTH.

2.5.1 TRAINING PHASE

While devising our approach, we perceived that one crucial requirement is that all resources and predicates in the source and target KGs must have at least one label via a

common predicate such as `rdfs:label`.[28] This avoids the generation of inadequate resources. After establishing that, we divide THOTH into two models in order to take into account the challenge of translating datatype property values (i.e., texts) and object property values (i.e., entities). Trying to tackle both kinds of statements with a single model is likely to fail as labels can easily reach a length of 50 characters. Therefore, we divide the data gathering process into two blocks in order to be able to train two models.

Data gathering process - First, we upload the source and target KG into a SPARQL endpoint and query both graphs by looking for resources which have the same "identity". Identical resources are usually connected via `owl:sameAs` links. However, aligned triples must not contain `owl:sameAs` as predicates in themselves. Second, we perform another SPARQL query for gathering only the labels of the aligned resources. Thus, we generate two bilingual training files, one with triples and another with labels (see Listing 2.1 for an example). Once both training files are created, we split them into training, development, and test sets.

```
EN: dbr:crocodile_dundee_ii dbo:country dbr:united_states
DE: dbr_de:crocodile_dundee_ii dbo:country dbr_de:vereinigte_staaten
EN: dbr:til_there_was_you dbo:writer dbr:winnie_holzman
DE: dbr_de:zwei_singles_in_l.a. dbo:writer dbr_de:winnie_holzman
```

Listing 2.1: Sample of the triple-based training data

Preprocessing - Before we start training the triple- and text-based models, we tokenize both training data files. Subsequently, we apply BPE models on them for dealing with OOV words (Sennrich et al., 2016a). BPE is a form of data compression that iteratively replaces the most frequent pair of bytes in a sequence with a single, unused byte. Applying BPE on the training data allows the translation models to translate words and sub-words and consequently improve their translation performance.

Knowledge Graph Embeddings - Based on recent findings (Moussallem et al., 2019a), we generate KGEs from the aligned triples along with their labels by using *fastText*. We rely on multinomial logistic regression (Böhning, 1992) as a classifier in a supervised training implemented in *fastText*. It assigns the entity's URI to its surface forms. This technique enables the NN to retrieve from KGE the surface form of the entities through their URIs.

[28] https://www.w3.org/TR/webont-req/section-requirements

Training - Both triple- and text-based models rely on a standard RNN model. The difference between both models is the training data format. The Triple-based model is trained only with the aligned triples, while the text-based was trained with an external generic bilingual corpora. Additionally, both models are augmented with the same KGE model. The idea of using KGE is to maximize the vector values of the triple-based and text-based NMT embeddings layers while training their models. An overview of the training phase can be found in Figure 2.8.

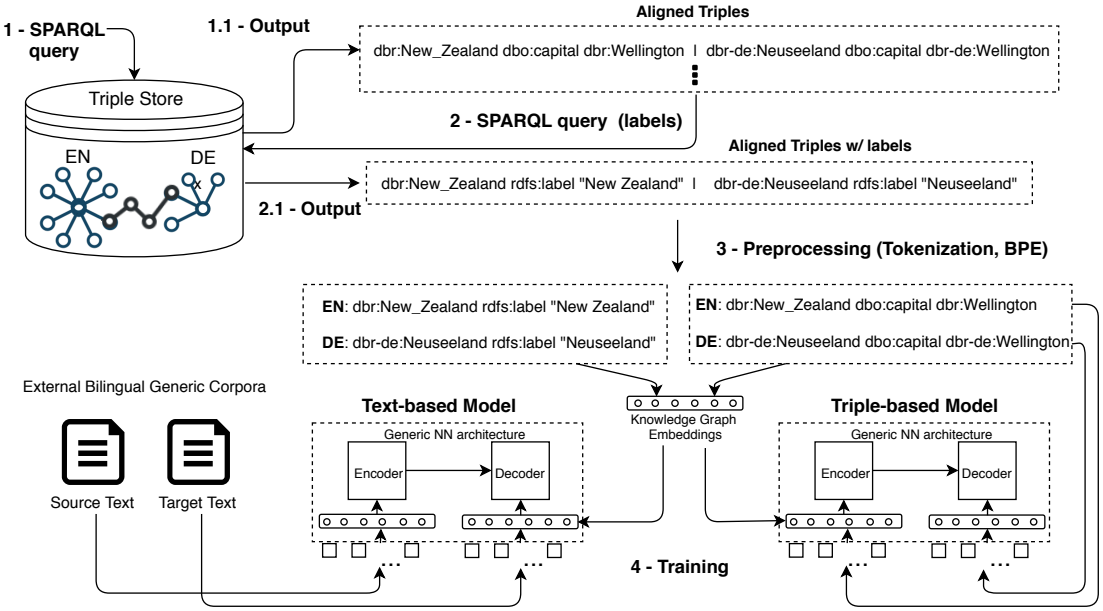

Figure 2.8: Training phase overview

2.5.2 TRANSLATION PHASE

Here, THOTH expects the entire source KG as an input to be translated and enriched into the target language as an output. To this end, THOTH first relies on a script, which is responsible for splitting the KG triples that comprises only the resources in one file and the triples that contain literals as objects in a different file. Once the division is done, and two set files are generated, THOTH starts translating the triples only with resources. After that, THOTH has to deal with the triples which have labels, and such triples are handled differently. The subject and predicate of the triples are sent to the Triple-based model along with a special character in the place of its object. This special character

simply tells the model to ignore the value and copy it to the target. In turn, the Text-based NMT model translates only the object. We argue that the Text-based model can translate the labels correctly since its model was augmented with a KGE model representing the URIs of both KGs, source and target. Afterwards, subject and predicate are attached with their object literal in a triple again. Finally, the two different files are combined into one again resulting in a translated KG. An overview of the training phase can be found in Figure 2.9.

Once the translation step is complete, THOTH gets the translated KG, and the original target (German) KG used in the training part and combines both into a single KG. The idea here is to enrich the original KG with translated triples. When conflicts of values happen, e.g., the triples match partially, and duplicated triples appear between the original KG and the translated KG, we opt to maintain the triples from the original KG as THOTH's aim is not to produce a newly translated KG but enrich the original one.

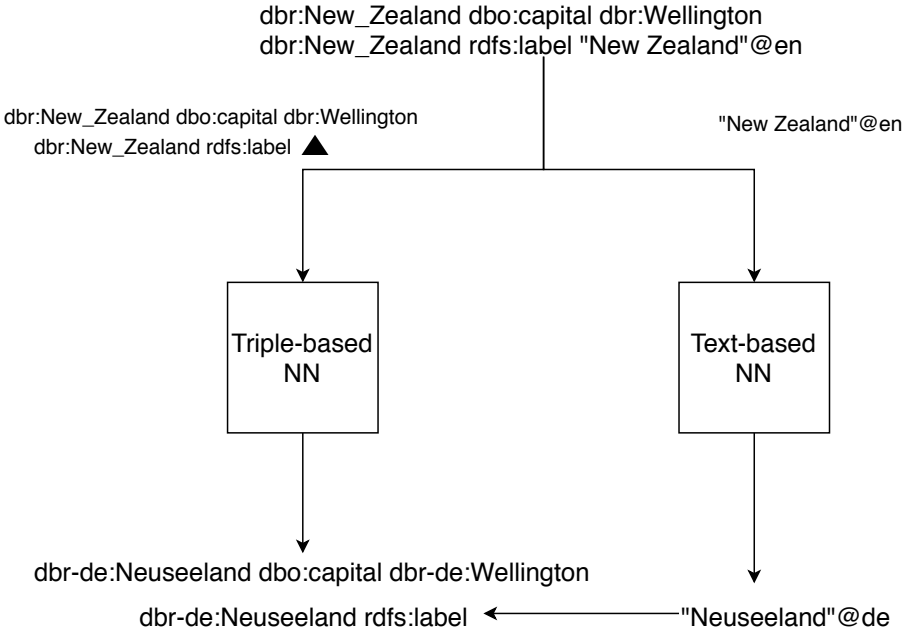

Figure 2.9: Translation phase overview

2.5.3 EVALUATION

We designed our evaluation in three-fold set. First, we measured the performance of THOTH using an automatic MT evaluation metric, BLEU, along with its translation

accuracy. Second, we evaluated THOTH extrinsically by comparing the German DBpedia with the German translation of the English DBpedia on two tasks: Fact Validation and Entity Linking. Third, we ran a manual intrinsic evaluation of the translation. We choose German as a target language because of the abundance of benchmarking systems and datasets for this pair.

Settings - In our experiments, both the triple-based and the text-based NMT models are built upon an RNN architecture using a bi-directional 2-layer LSTM encoder-decoder model with attention mechanism (Bahdanau et al., 2014). The training uses a batch size of 32 and the stochastic gradient descent with an initial learning rate of 0.0002. We set the dimension of the word embeddings to 500 and the internal embeddings of hidden layers to size 500. The dropout is set to 0.3 (naive). We use a maximum sentence length of 50, a vocabulary of 50,000 words and a beam size of 5. All experiments are performed with the OpenNMT framework (Klein et al., 2017). In addition, we encode the triples and words using BPE (Sennrich et al., 2016a) with 32,000 merge operations. For training the text-based model, our training set consists of a merge of all parallel training data provided by the Workshop on Machine Translation (WMT) tasks[29], obtaining after preprocessing a corpus of five million sentences with 79M running words. In the triple-based model, we use the bilingual alignments from the English, and German versions of DBpedia[30] for training. This alignment contains 346,373 subjects, 292 relations and 208,079 objects in 1,012,681 triples. We divide this data into 80% training, 10% development and 10% test. Overall, the English KG contains 4.2 million entities, 661 relations, and 2.1 million surface forms, while the German version has 1 million entities, 249 relations, and 0.5 million surface forms. Additionally, we train the KGE on both DBpedia versions using the *fastText* algorithm with a vector dimension size of 500 and a window size of 50 by using 12 threads with hierarchical softmax.

Translation task - The overall enrichment quality of THOTH is measured by working through different steps. Firstly, we evaluate the translations automatically by computing a translation accuracy with BLEU (Papineni et al., 2002b) score. In the subsequent evaluation steps, we investigate THOTH's performance on a full KG translation setting. In this case, we use THOTH models for translating and enriching all CBD resources

[29] http://www.statmt.org/wmt18/translation-task.html

[30] We selected the subsets of mapping-based objects and labels to evaluate the quality of our approach since they are the most used ones for training Linked-Data NLP approaches.

of English DBpedia to an enriched-German DBpedia version. The further extrinsic evaluation steps are described below.

Fact validation task - We selected FactBench—a multilingual benchmark dataset for the evaluation of fact validation algorithms (Gerber et al., 2015)—for our experiments. FactBench contains positive and negative facts. We only use the 750 positive facts distributed over 10 relations as reference data in our experiment. Our aim is to check the number of true facts which existed in the original KG (i.e., in the German version of DBpedia) and how many true triples THOTH was able to add to the KG through enrichment. We used 5 of the 10 predicates in our evaluation data set, i.e., `award`, `birthplace`, `deathplace`, `leader`, `starring` because the other predicates do not lead to sufficient training data. Overall, our evaluation dataset consists of a total count of 375 facts.

NLP task - Our idea here is to exploit the graphs connections from the enriched-German DBpedia (THOTH) KG to improve a given EL system on a disambiguation task. We chose MAG, our multilingual EL system introduced by Moussallem et al. (2017), which is language- and KG-agnostic. MAG does not require any training even though shows competitive results. Also, we selected GERBIL (Usbeck et al., 2015) as a benchmarking platform. As the evaluation is on the German language, we uploaded four German datasets to GERBIL.

2.5.4 RESULTS

In this section, we report the results of THOTH's enrichment in the German DBpedia on the settings mentioned above.

Translation results - We evaluated our translation on the test set of the bilingual data we extracted via SPARQL queries. THOTH achieved a BLEU score of 65.47, which is superior to the state-of-the-art translation scores achieved on natural language (Edunov et al., 2018).

Given that it is not possible to infer the quality of a given translation only relying on one automatic evaluation metric, we created an additional evaluation script that computes the exact string match of subjects, predicates, and objects between an output and a reference translation triple. Additionally, we also computed the overall triple accuracy. Figure 2.10 depicts the accuracy results of THOTH's output in comparison to the German test set. THOTH achieved up to 80% accuracy for subjects, predicates, and objects. As

expected, THOTH's accuracy decreased to 68.83% when measuring entire triples. We analyzed the results manually to understand this drop in the performance. Our manual analysis suggests that the poorer performance w.r.t. triples is linked to the partially weak disambiguation power of the underlying KGE model, which assigned the same vector value for similar predicates. Our results confirm that NNs along KGE can support a full KG translation by considering the consistent quality of THOTH translations.

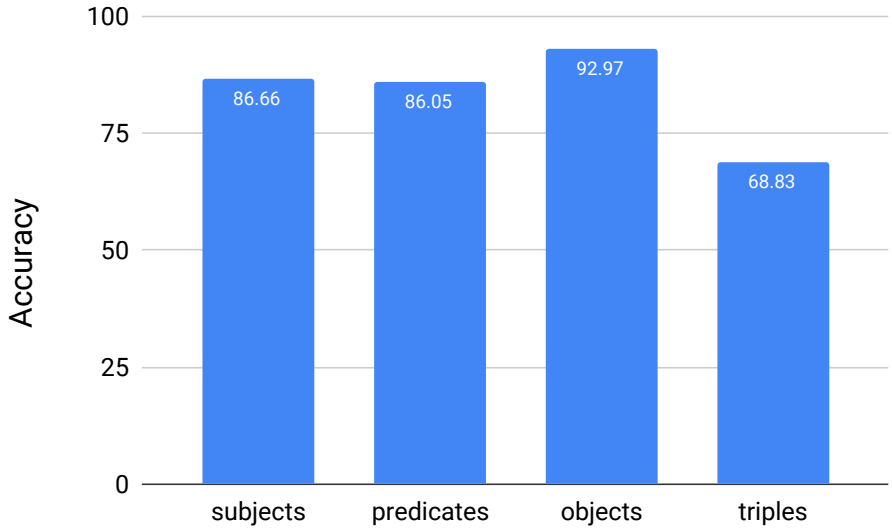

Figure 2.10: Overall translation accuracy

Fact validation results - Here, we used THOTH to translate the entire English DBpedia to German. In this case, we do not have a gold standard translation to compare automatically. Therefore, we evaluated the THOTH's enrichment capability in the perspective of a fact-validation task. The main goal here was to check if THOTH could enrich the original German KG with new correct facts which were not present in its original version. Figure 2.11 reports an improvement of 18.4% across all predicates. THOTH led to a significant increase in the number of correct facts in the original KG.

Entity Linking results - For this evaluation, we used the optimal parameter configuration for MAG described by Moussallem et al. (2017). Table 2.11 reports the results of MAG in two configuration sets, one with original German DBpedia and another with the *Enriched-German DBpedia (THOTH)* as reference KGs. The version of MAG running on the translated KG achieves significantly better results than that running on the original KG. The average improvement across all datasets is around 19% in F-measure.

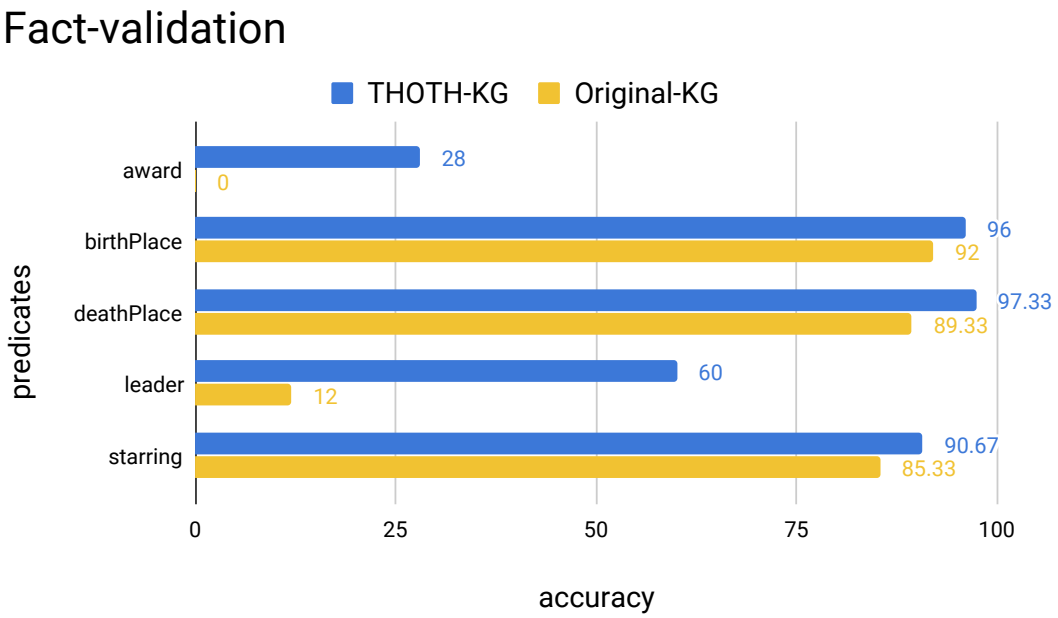

Figure 2.11: A comparison between the enriched-German DBpedia (THOTH) KG with the original German DBpedia on the validation of facts.

The results of the *German abstracts* data set and N^3 *news.de* are surprisingly high. We sampled the results manually, and we established that the results were correct. We also investigated the creation of both benchmarking datasets, and concluded that at the time of their creation, the links used in both were based on the English DBpedia as an auxiliary KG. Therefore, when THOTH translated the English KG to German and enriched the original German DBpedia with English knowledge, MAG was able to get very high HITS scores for many resources.

Table 2.11: Micro results in a comparison between German DBpedia KG with Enriched-German DBpedia (THOTH) KG in MAG.

Datasets	**MAG-DBpedia-KG**			**MAG-THOTH-KG**		
	F-measure	Precision	Recall	F-measure	Precision	Recall
German Abstracts	0.78	0.79	0.76	**0.97**	**0.99**	**0.96**
N^3 news.de	0.77	0.78	0.76	**0.98**	**0.99**	**0.97**
VoxEL-strict	0.40	0.46	0.35	**0.70**	**0.81**	**0.61**
VoxEL-relaxed	0.57	0.57	0.57	**0.64**	**0.64**	**0.64**

2.5.5 REPRODUCIBILITY

all experimental data, code, models are publicly available.[31]

2.5.6 SUMMARY

The main contributions of this paper can be summarized as follows:

- We present a novel approach based on NNs along with KGEs for translating and enriching KGs across languages.
- THOTH achieves a translation accuracy of 88.56% across all elements of a triple.
- THOTH improves the quality of the original German DBpedia significantly in both the fact checking and the EL tasks: 18.4% for fact validation and 19% for EL.

[31] https://github.com/dice-group/THOTH

3
Conclusions and Outlook

This chapter concludes the thesis by summarizing our results in Section 3.1 and giving an outlook on future research directions in Section 3.2.

3.1 Conclusions

Handling entities across distinct NLP tasks is a difficult task. With thesis, we showed that real-world KGs can contribute to improve the results of EL, MT and NLG tasks. Therewith, we answered our main research question posed in Section 1.1:

RQ. Can KGs alleviate the ambiguity problem and be used to improve the quality of automatic text translation and generation?

In addition, we addressed its respective challenges by (1) devising a multilingual KG-based EL approach; (2) developing a multilingual RDF-to-Text verbalizer; (3) creating the first neural- and KG-based REG model; (4) creating the first KG-augmented NMT model; (5) designing the first neural translation approach for enriching low resource KGs.

3.1.1 MAG: A Multilingual, Knowledge-base Agnostic and Deterministic Entity Linking Approach

We asked the following two questions in Section 1.1.1:

RQ1. Can a KG-based EL approach achieve a similar performance across languages?

RQ2. Does a language-based KG influence the disambiguation quality?

We answered the above mentioned questions by presenting MAG, a KB-agnostic and deterministic approach for multilingual EL. MAG outperforms state of the art on all non-English data sets. In addition, MAG achieves a performance similar to state of the art on English data sets. An average 0.63 F-measure places MAG 1st out of 13 annotation systems. Furthermore, we analyzed the influence of different indexing and searching methods, as well as the influence of the data set structure in a fine-grained evaluation. We also provided a context search without relying on machine learning, as previously done. Moreover, we showed that current ML-based EL approaches are strongly biased due to their learned model. This behavior can be seen in multilingual data sets. We also deployed and analyzed the influence of acronyms and last names.

3.1.2 RDF2PT: GENERATING BRAZILIAN PORTUGUESE TEXTS FROM RDF DATA

We asked the following question in Section 1.1.2:

RQ3: Can KGs as input support the generation of multilingual text?

We answered the above mentioned question by presenting RDF2PT, the first approach that verbalizes RDF data to Brazilian Portuguese texts. Compared with human texts, RDF2PT generates texts with high fluency and clarity. We identified essential challenges for generating multilingual texts from RDF using a rule- and template-based approach. Moreover, we extended RDF2PT to Spanish and English (Ngonga Ngomo et al., 2018; Ngonga Ngomo et al., 2019), thus demonstrating that KGs definitely support the multilingualism in NLG.

3.1.3 NEURALREG: AN END-TO-END APPROACH TO REFERRING EXPRESSION GENERATION

We asked the following question in Section 1.1.2:

RQ4: Can KGs be used for accomplishing the full REG task?

We answered the above mentioned question by introducing NeuralREG, the first end-to-end approach based on neural networks for the REG task. NeuralREG generates referring expressions for discourse entities by simultaneously selecting form and content without any need for feature extraction techniques. NeuralREG showed that the neural model substantially improves over two strong baselines, both in terms of the accuracy of referring expressions and the fluency of lexicalized texts.

3.1.4 KG-NMT: UTILIZING KNOWLEDGE GRAPHS FOR NEURAL MACHINE TRANSLATION AUGMENTATION

We asked the following question in Section 1.1.3:

RQ5: Can an NMT model enhanced with a bilingual KG improve translation quality?

We answered the above mentioned question by presenting KG-NMT. KG-NMT is the first augmentation methodology, which relies on the use of KGs to improve the performance of NMT systems for translating domain-specific expressions and named entities in texts. We implemented two strategies for incorporating KGEs into NMT models that work on word- and sub-word units-based models. Additionally, we carried out an extensive evaluation with a manual analysis, which showed consistent translation improvements provided by incorporating DBpedia KG in NMT. The overall methodology can be applied to any NMT model since it does not modify the main NMT model structure and also allows the replacement of different EL systems.

3.1.5 THOTH: TRANSLATING AND ENRICHING LOW-RESOURCE KG

We asked the following two questions in Section 1.1.4:

RQ6: Can NMT support a full (triples and labels) translation of KGs?

RQ7: Can an artificially-enriched KG improve the performance of a system on NLP tasks?

We answered the above mentioned questions by introducing THOTH, the first neural-based approach for translating and enriching KGs from different languages. THOTH is a promising approach that achieves a translation accuracy of 88.56%. Moreover, its enrichment improves the quality of the German DBpedia significantly, as we report

+18.4% accuracy for fact validation and +19% F_1 for entity linking. THOTH relies on two different RNN-based NMT models along with KGEs for translating triples and texts jointly. We carried out an extensive evaluation set for certifying the quality of our approach.

3.2 OUTLOOK

Extensions of our contributions could be performed in multiple directions: exploiting other KG features, such as ontologies within NMT models, and extending our NLG approaches to other languages. Moreover, a further investigation of our KG translation and enrichment approach on other NN architectures along other KGE algorithms has to be carried out.

3.2.1 EXPLOITING KG FEATURES

The syntactic disambiguation problem still lacks good solutions. For instance, the English language contains irregular verbs like “set” or “put”. Depending on the structure of a sentence, it is not possible to recognize their verbal tense, e.g., present or past tense. Even statistical approaches trained on huge corpora may fail to find the exact meaning of some words due to the structure of the language. Although this challenge has successfully been dealt with since NMT has been used for European languages (Bojar et al., 2017), implementations of NMT for some non-European languages have not been fully exploited (e.g., Brazilian Portuguese, Latin-America Spanish, Hindi) due to the lack of large bilingual data sets on the Web to be trained on. We suggest using ontology properties via semantic annotations to alleviate the syntactic issue of irregular verbs. For instance, the sentence “Anna usually put her notebook on the table for studying" may be annotated using a given vocabulary by triples. Thus, the verb “put", which is represented by a predicate that groups essential information about the verbal tense, may support the generation step of a given NMT system. This sentence usually fails when translated to morphologically rich languages, such as Brazilian-Portuguese and Arabic, for which the verb influences the translation of “usually" to the past tense. In this case, the ontology properties contained in a KG may support the problem of finding a specific rule behind relationships between source and target texts in the training phase. Some researchers, including Harriehausen-Mühlbauer and Heuss (2012); Seo et al. (2009),

have used ontology properties to disambiguate words in MT systems. However, the ontologies were not exploited in the context of NMT.

To include ontology properties as features in NMT models, some steps need to be addressed. Currently, the Ontology-Lexica Community Group[1] at W3C has combined efforts to represent lexical entries, with their linguistic information, in ontologies across languages. Modeling different languages using the same model may provide alignment between the languages, where it is possible to infer new rules using the language dependency graph structure and visualize a similarity among languages. [2]

3.2.2 COMBINING THE MODELS OF NMT WITH NLG

During the course of this thesis, we realized that combining a KG-augmented NMT model with an REG model can be fruitful for improving the fluency of entities in text translation. The idea relies on feeding the NeuralREG model with the output of KG-NMT, but instead of generating the final translation with KG-NMT, it keeps the URI of the entities in the NMT output. Figure 3.1 depicts the general idea of this promising direction.

3.2.3 DEEPER INVESTIGATION OF MULTILINGUALISM IN NLG

During the development of RDF2PT, some challenges became clear while handling morphologically rich languages. For example, recognizing gender continues to be a hard task. For example, in "`Os Lusíadas é uns obra literária`", the determiner `uns` should be feminine and singular, because `obra` is singular and has a feminine gender. However, it is accorded to the subject `Os Lusíadas`. Although our approach, NeuraLREG, was capable of improving the generation of pronouns, it still requires improvements and investigation with regard to other languages. We hence envisage adding gender as a feature for improving the text generation.

Another observed challenge was the generation of coordinated sentences by RDF2PT and NeuralREG, which helped the users in our experimental setup recognize if the models or humans generated a given text. This behavior arises because humans are likely to write subordinate sentences. For example, while RDF2PT can generate `Albert Einstein`

[1] https://www.w3.org/community/ontolex/

[2] This insight is already supported by a recent publication at the Cicling Conference https://www.cicling.org/2017/posters.html named "The Fix-point of Dependency Graph – A Case Study of Chinese-German Similarity" by Tiansi Dong et al.

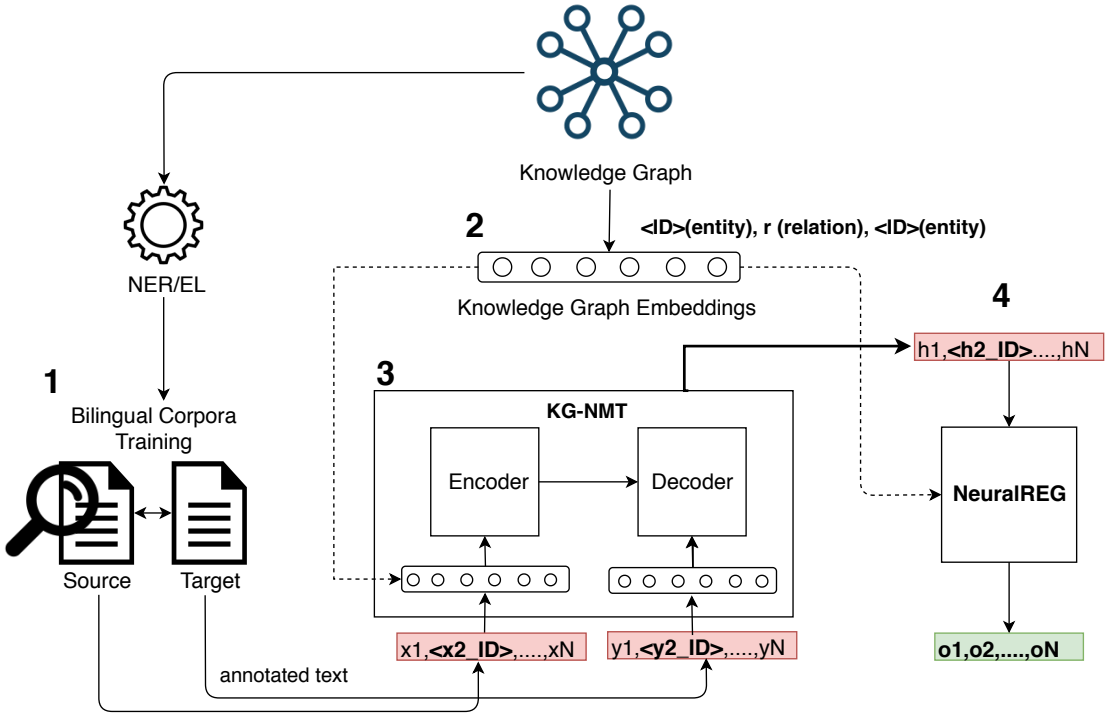

Figure 3.1: Overview of the combination of KG-NMT + NeuralREG.

`foi um cientista e ele nasceu em Ulm.` (eng: Albert Einstein was a scientist and he was born in Ulm), a human would write this same sentence in the following way, `Albert Einstein foi um cientista que/cujo nasceu em Ulm` (eng: Albert Einstein was a scientist who was born in Ulm). This difference was crucial in the perspective of our evaluators. Therefore, the generation of subordinate sentences must be investigated.

3.2.4 TRANSLATING KGS WITH OTHER NNS

While analyzing THOTH's output, we noticed some mistranslations of similar predicates that were responsible for decreasing the accuracy of the triple translation. For example, the following English source triple `dbr:zenyattà_mondatta dbo:artist dbr:the_police` was translated into `dbr_de:zenyattà_mondatta dbo:producer dbr_de:the_police`. This example shows that THOTH translated the subject and object correctly. However, the predicate was incorrect and was mistranslated from dbo:artist to dbo:producer. A similar problem occurred while translating

the triple, `dbr:albert_einstein dbo:citizenship dbr:Switzerland` to `dbr:albert_einstein dbo:birthplace dbr:der_Schweiz`. After a manual analysis, we identified that both cases happened because THOTH could not distinguish the predicates that share the same domain and range. In a more in-depth analysis, we perceived that the predicates mentioned above are very close to each other in the vector space, thus complicating the disambiguation process of NN models. The performance of THOTH was not affected by these false triples since they were automatically removed in the enrichment step. After this manual analysis of the results, we believe that addressing the problem of similar predicates (e.g., through novel embedding techniques) can enhance the translation quality of THOTH. The application of sub-graphs (Cao et al., 2018) and other NN architectures, such as Transformer (Vaswani et al., 2017), for improving the disambiguation of similar predicates are promising paths.

References

Aprosio, A. P., C. Giuliano, and A. Lavelli. Towards an automatic creation of localized versions of DBpedia. In *International Semantic Web Conference*, pages 494–509. Springer, 2013.

Arnold, D. *Machine translation: an introductory guide*. Blackwell Pub, 1994.

Auer, S., C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer, 2007.

Bahdanau, D., K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Banerjee, S. and A. Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 65–72. ACL, 2005.

Bar-Hillel, Y. The present status of automatic translation of languages. In *Advances in computers*, volume 1, pages 91–163. Elsevier, 1960.

Belinkov, Y. and J. Glass. Analysis Methods in Neural Language Processing: A Survey. *Transactions of the Association for Computational Linguistics*, 7:49–72, March 2019. doi: 10.1162/tacl_a_00254. URL https://www.aclweb.org/anthology/Q19-1004.

Berners-Lee, T., J. Hendler, and O. Lassila. The Semantic Web. *Scientific american*, 284 (5):34–43, 2001.

Bisazza, A. and M. Federico. A survey of word reordering in statistical machine translation: Computational models and language phenomena. *Computational Linguistics*, 2016.

Böhning, D. Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics*, 1:197–200, 1992.

Bojar, O., R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, S. Huang, M. Huck, P. Koehn, Q. Liu, V. Logacheva, et al. Findings of the 2017 conference on machine translation (WMT17). In *Proceedings of the Second Conference on Machine Translation*, pages 169–214, 2017.

Bonatti, P. A., S. Decker, A. Polleres, and V. Presutti. Knowledge graphs: new directions for knowledge representation on the semantic web (Dagstuhl seminar 18371). *Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik*, 2019.

Brown, P. F., J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85, 1990.

Bryl, V., C. Bizer, and H. Paulheim. Gathering alternative surface forms for DBpedia entities. In *NLP & DBpedia 2015*, volume 1581, pages 13–24, Aachen, 2016. RWTH.

Callaway, C. B. and J. C. Lester. Pronominalization in Generated Discourse and Dialogue. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL'02, pages 88–95, Philadelphia, Pennsylvania, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073100.

Cao, Z., L. Wang, and G. de Melo. Link Prediction via Subgraph Embedding-Based Convex Matrix Completion. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

Carpuat, M. and D. Wu. Improving Statistical Machine Translation Using Word Sense Disambiguation. In *EMNLP-CoNLL*, volume 7, pages 61–72, 2007.

Castro Ferreira, T., E. Krahmer, and S. Wubben. Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, ACL'16, pages 568—-577, Berlin, Germany, 2016. Association for Computational Linguistics.

Castro Ferreira, T., E. Krahmer, and S. Wubben. Generating flexible proper name references in text: Data, models and evaluation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, EACL'17, pages 655–664, Valencia, Spain, 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/E17-1062.

Chung, J., K. Cho, and Y. Bengio. A Character-level Decoder without Explicit Segmentation for Neural Machine Translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1693–1703. ACL, 2016.

Clark, J. H., C. Dyer, A. Lavie, and N. A. Smith. Better Hypothesis Testing for Statistical Machine Translation: Controlling for Optimizer Instability. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, ACL'11, pages 176–181, Portland,

Oregon, 2011. ISBN 978-1-932432-88-6. URL http://dl.acm.org/citation.cfm?id=2002736.2002774.

Colin, E., C. Gardent, Y. Mrabet, S. Narayan, and L. Perez-Beltrachini. The webnlg challenge: Generating text from dbpedia data. In *Proceedings of the 9th International Natural Language Generation conference*, pages 163–167, 2016.

Costa-jussà, M. R. How much hybridization does machine translation Need? *Journal of the Association for Information Science and Technology*, 66(10):2160–2165, 2015.

Costa-Jussà, M. R. and M. Farrús. Statistical machine translation enhancements through linguistic levels: A survey. *ACM Computing Surveys (CSUR)*, 46(3):42, 2014.

Costa-Jussa, M. R. and J. A. Fonollosa. Latest trends in hybrid machine translation and its applications. *Computer Speech & Language*, 32(1):3–10, 2015.

Costa-Jussa, M. R., M. Farrús, J. B. Mariño, and J. A. Fonollosa. Study and comparison of rule-based and statistical Catalan-Spanish machine translation systems. *Computing and Informatics*, 31(2):245–270, 2012.

Dale, R. and N. Haddock. Generating referring expressions involving relations. In *Proceedings of the fifth conference on European chapter of the Association for Computational Linguistics*, EACL'91, pages 161–166, Berlin, Germany, 1991. Association for Computational Linguistics. doi: 10.3115/977180.977208.

Dale, R. and E. Reiter. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive science*, 19(2):233–263, 1995.

De Oliveira, R. and S. Sripada. Adapting SimpleNLG for Brazilian Portuguese realisation. In *INLG*, pages 93–94, 2014.

Devi, P., A. Gupta, and A. Dixit. Comparative Study of HITS and PageRank Link based Ranking Algorithms. *International Journal of Advanced Research in Computer and Communication Engineering*, 3(2):5749–5754, 2014.

Edunov, S., M. Ott, M. Auli, and D. Grangier. Understanding Back-Translation at Scale. *arXiv preprint arXiv:1808.09381*, 2018.

Ell, B., D. Vrandecic, and E. P. B. Simperl. Labels in the Web of Data. In *Proceedings of ISWC*, volume 7031, pages 162–176. Springer, 2011.

Ferreira, T. C., E. Krahmer, and S. Wubben. Towards more variation in text generation: Developing and evaluating variation models for choice of referential form. In *ACL (1)*, 2016.

Ferreira, T. C., E. Krahmer, and S. Wubben. Generating flexible proper name references in text: Data, models and evaluation. In *Proc. EACL*, volume 17, 2017.

Ferreira, T. C., D. Moussallem, E. Krahmer, and S. Wubben. Enriching the WebNLG corpus. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 171–176, 2018a.

Ferreira, T. C., D. Moussallem, Ákos Kádár, S. Wubben, and E. Krahmer. NeuralREG: An end-to-end approach to referring expression generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018b.

Friedman, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200): 675–701, 1937.

Ganea, O.-E., M. Ganea, A. Lucchi, C. Eickhoff, and T. Hofmann. Probabilistic Bag-Of-Hyperlinks Model for Entity Linking. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 927–938, Republic and Canton of Geneva, Switzerland, 2016. International World Wide Web Conferences Steering Committee. ISBN 978-1-4503-4143-1. doi: 10.1145/2872427.2882988. URL http://dx.doi.org/10.1145/2872427.2882988.

Gardent, C., A. Shimorina, S. Narayan, and L. Perez-Beltrachini. Creating Training Corpora for NLG Micro-Planners. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL'17, pages 179–188, Vancouver, Canada, 2017a. Association for Computational Linguistics. doi: 10.18653/v1/P17-1017. URL http://www.aclweb.org/anthology/P17-1017.

Gardent, C., A. Shimorina, S. Narayan, and L. Perez-Beltrachini. The WebNLG Challenge: Generating Text from RDF Data. In *Proceedings of the 10th International Conference on Natural Language Generation*, INLG'17, pages 124–133, Santiago de Compostela, Spain, 2017b. Association for Computational Linguistics. URL http://aclweb.org/anthology/W17-3518.

Gardent, C., A. Shimorina, S. Narayan, and L. Perez-Beltrachini. Creating training corpora for nlg micro-planning. In *Proceedings of ACL*, 2017c.

Gatt, A. and E. Krahmer. Survey of the State of the Art in Natural Language Generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170, 2018.

Gerber, D., D. Esteves, J. Lehmann, L. Bühmann, R. Usbeck, A.-C. N. Ngomo, and R. Speck. Defacto—temporal and multilingual deep fact validation. *Web Semantics: Science, Services and Agents on the World Wide Web*, 35:85–101, 2015.

Glorot, X. and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR. URL http://proceedings.mlr.press/v9/glorot10a.html.

Grave, E., P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2018.

Harriehausen-Mühlbauer, B. and T. Heuss. Semantic web based machine translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pages 1–9. Association for Computational Linguistics, 2012.

Henschel, R., H. Cheng, and M. Poesio. Pronominalization Revisited. In *Proceedings of the 18th Conference on Computational Linguistics - Volume 1*, COLING'00, pages 306–312, Saarbrücken, Germany, 2000. Association for Computational Linguistics. ISBN 1-55860-717-X. doi: 10.3115/990820.990865. URL https://doi.org/10.3115/990820.990865.

Heuss, T. Lessons learned (and questions raised) from an interdisciplinary Machine Translation approach. In *Position paper for the W3C Workshop on the Open Data on the Web*, pages 23–24, 2013.

Hochreiter, S. and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Hoffart, J., M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust Disambiguation of Named Entities in Text. In *Conference on Empirical Methods in Natural Language Processing*, 2011.

Hoffart, J., Y. Altun, and G. Weikum. Discovering Emerging Entities with Ambiguous Names. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 385–396, New York, NY, USA, 2014. ACM.

Hutchins, W. J. and H. L. Somers. *An introduction to machine translation*, volume 362. Academic Press London, 1992.

Joulin, A., E. Grave, P. Bojanowski, M. Nickel, and T. Mikolov. Fast Linear Model for Knowledge Graph Embeddings. *arXiv preprint arXiv:1710.10881*, 2017.

Jurafsky, D. *Speech and language processing: An introduction to natural language processing*. Prentice Hall, 2000.

K M, A., S. Basu Roy Chowdhury, and A. Dukkipati. Learning beyond Datasets: Knowledge Graph Augmented Neural Networks for Natural Language Processing. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 313–322. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/N18-1029.

Keet, C. M. and L. Khumalo. Toward a knowledge-to-text controlled natural language of isiZulu. *Language Resources and Evaluation*, 51(1):131–157, 2017.

Klein, G., Y. Kim, Y. Deng, J. Senellart, and A. M. Rush. OpenNMT: Open-Source Toolkit for Neural Machine Translation. *ArXiv e-prints*, 2017.

Kleinberg, J. M. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5): 604–632, 1999.

Koehn, P. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005.

Koehn, P. *Statistical Machine Translation*. Cambridge University Press, 2010.

Koehn, P. and R. Knowles. Six Challenges for Neural Machine Translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, 2017.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL*, pages 177–180. Association for Computational Linguistics, 2007.

Krahmer, E. and K. Van Deemter. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218, 2012a.

Krahmer, E. and K. Van Deemter. Computational generation of referring expressions: A survey. *Computational Linguistics*, 38(1):173–218, 2012b.

Lakshen, G. A., V. Janev, and S. Vraneš. Challenges in Quality Assessment of Arabic DBpedia. In *Proceedings of the 8th International Conference on Web Intelligence, Mining and Semantics*, page 15. ACM, 2018.

Levenshtein, V. I. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, February 1966.

Li, Z., X. Wang, A. Aw, E. S. Chng, and H. Li. Named-Entity Tagging and Domain adaptation for Better Customized Translation. In *Proceedings of the Seventh Named Entities Workshop*, pages 41–46. ACL, 2018.

Libovický, J. and J. Helcl. Attention Strategies for Multi-Source Sequence-to-Sequence Learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL'17, pages 196–202, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2031. URL http://www.aclweb.org/anthology/P17-2031.

Lopez, A. and M. Post. Beyond bitext: Five open problems in machine translation. In *Proceedings of the EMNLP Workshop on Twenty Years of Bitext*, pages 1–3, 2013.

Luong, M.-T. and C. D. Manning. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 76–79, 2015.

Luong, M.-T. and C. D. Manning. Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1054–1063. ACL, 2016.

Moussallem, D., R. Usbeck, M. Röeder, and A.-C. N. Ngomo. MAG: A Multilingual, Knowledge-base Agnostic and Deterministic Entity Linking Approach. In *Proceedings of the Knowledge Capture Conference*, page 9. ACM, 2017.

Moussallem, D., T. C. Ferreira, M. Zampieri, M. C. Cavalcanti, G. Xexéo, M. Neves, and A.-C. N. Ngomo. RDF2PT: Generating Brazilian Portuguese Texts from RDF Data. In *The 11th edition of the Language Resources and Evaluation Conference, 7-12 May 2018, Miyazaki (Japan)*, 2018a. URL https://arxiv.org/abs/1802.08150.

Moussallem, D., R. Usbeck, M. Röder, and A.-C. N. Ngomo. Entity Linking in 40 Languages using MAG. In *The Semantic Web, ESWC 2018, Lecture Notes in Computer Science*, 2018b.

Moussallem, D., M. Wauer, and A.-C. N. Ngomo. Machine Translation Using Semantic Web Technologies: A Survey. *Journal of Web Semantics*, 51:1–19, 2018c.

Moussallem, D., A.-C. N. Ngomo, P. Buitelaar, and M. Arcan. Utilizing Knowledge Graphs for Neural Machine Translation Augmentation. In *Proceedings of the 10th International Conference on Knowledge Capture*, pages 139–146. ACM, 2019a.

Moussallem, D., T. Soru, and A.-C. N. Ngomo. THOTH: Neural Translation and Enrichment of Knowledge Graphs. In *The Semantic Web ISWC 2019*, pages 1–17. Springer, 2019b.

Navigli, R. Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41 (2):10, 2009.

Neishi, M., J. Sakuma, S. Tohda, S. Ishiwatari, N. Yoshinaga, and M. Toyoda. A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size. In *Proceedings of the 4th Workshop on Asian Translation*, pages 99–109, 2017.

Neubig, G., C. Dyer, Y. Goldberg, A. Matthews, W. Ammar, A. Anastasopoulos, M. Ballesteros, D. Chiang, D. Clothiaux, T. Cohn, K. Duh, M. Faruqui, C. Gan, D. Garrette, Y. Ji, L. Kong, A. Kuncoro, G. Kumar, C. Malaviya, P. Michel, Y. Oda, M. Richardson, N. Saphra, S. Swayamdipta, and P. Yin. DyNet: The Dynamic Neural Network Toolkit. *ArXiv e-prints*, January 2017.

Ngonga Ngomo, A.-C., L. Bühmann, C. Unger, J. Lehmann, and D. Gerber. Sorry, i don't speak SPARQL: translating SPARQL queries into natural language. In *Proceedings of the 22nd international conference on World Wide Web*, pages 977–988. ACM, 2013.

Ngonga Ngomo, A.-C., M. Röder, D. Moussallem, R. Usbeck, and R. Speck. BENGAL: An Automatic Benchmark Generator for Entity Recognition and Linking. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 339–349, 2018.

Ngonga Ngomo, A.-C., D. Moussallem, and L. Bühman. A Holistic Natural Language Generation Framework for the Semantic Web. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, page 8. ACL (Association for Computational Linguistics), 2019.

Orita, N., E. Vornov, N. Feldman, and H. Daumé III. Why discourse affects speakers' choice of referring expressions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, ACL'15, pages 1639–1649, Beijing, China, 2015. Association for Computational Linguistics. doi: 10.3115/v1/P15-1158. URL http://www.aclweb.org/anthology/P15-1158.

Page, L., S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, ACL'02, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002a. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL http://www.aclweb.org/anthology/P02-1040.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002b.

Popović, M. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, 2017.

Ramos, J. et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, 2003.

Reiter, E. and R. Dale. *Building natural language generation systems*. Cambridge University Press, New York, NY, USA, 2000. ISBN 0-521-62036-8.

Röder, M., R. Usbeck, and A.-C. N. Ngomo. GERBIL–Benchmarking Named Entity Recognition and Linking Consistently. *Semantic Web Journal*, 2018. URL http://www.semantic-web-journal.net/system/files/swj1671.pdf.

Schuster, M. and K. K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

Sennrich, R., B. Haddow, and A. Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1715–1725. ACL, 2016a.

Sennrich, R., B. Haddow, and A. Birch. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 86–96, 2016b.

Seo, E., I.-S. Song, S.-K. Kim, and H.-J. Choi. Syntactic and semantic English-Korean machine translation using ontology. In *Advanced Communication Technology, 2009. ICACT 2009. 11th International Conference on*, volume 3, pages 2129–2132. IEEE, 2009.

Siddharthan, A., A. Nenkova, and K. McKeown. Information Status Distinctions and Referring Expressions: An Empirical Study of References to People in News Summaries. *Computational Linguistics*, 37(4):811–842, 2011. doi: 10.1162/COLI_a_00077. URL http://dx.doi.org/10.1162/COLI_a_00077.

Slocum, J. A survey of machine translation: its history, current status, and future prospects. *Computational linguistics*, 11(1):1–17, 1985.

Sorokin, D. and I. Gurevych. Modeling Semantics with Gated Graph Neural Networks for Knowledge Base Question Answering. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3306–3317. ACL, 2018.

Stahlberg, F. Neural Machine Translation: A Review, 2019.

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufis, and D. Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058*, 2006.

Sutskever, I., O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

Thurmair, G. Comparing rule-based and statistical MT output. In *The Workshop Programme*, page 5, 2004.

Thurmair, G. Comparing different architectures of hybrid Machine Translation systems. *MT Summit XII: proceedings of the twelfth Machine Translation Summit*, pages 340–347, 2009.

Tiedemann, J. Parallel Data, Tools and Interfaces in OPUS. In Chair), N. C. C., K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, and S. Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, may 2012. European Language Resources Association (ELRA). ISBN 978-2-9517408-7-7.

Toutanova, K. and C. D. Manning. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics-Volume 13*, pages 63–70. Association for Computational Linguistics, 2000.

Ugawa, A., A. Tamura, T. Ninomiya, H. Takamura, and M. Okumura. Neural Machine Translation Incorporating Named Entity. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3240–3250, 2018.

Usbeck, R., A. N. Ngomo, M. Röder, D. Gerber, S. A. Coelho, S. Auer, and A. Both. AGDISTIS - Graph-Based Disambiguation of Named Entities Using Linked Data. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, October 19-23, 2014. Proceedings, Part I*, pages 457–471, Riva del Garda, Italy, 2014.

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

van Deemter, K. Designing Algorithms for Referring with Proper Names. In *Proceedings of the 9th International Natural Language Generation conference*, INLG'16, pages 31–35, Edinburgh, UK, 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-6605. URL http://www.aclweb.org/anthology/W16-6605.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Vrandečić, D. and M. Krötzsch. Wikidata: a free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014.

Waitelonis, J., H. Jürges, and H. Sack. Don'T Compare Apples to Oranges: Extending GERBIL for a Fine Grained NEL Evaluation. In *Proceedings of the 12th International Conference on Semantic Systems*, SEMANTiCS 2016, pages 65–72, New York, NY, USA, 2016. ACM.

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*, 2016.

Yang, B. and T. Mitchell. Leveraging knowledge bases in lstms for improving machine reading. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 1436–1446, 2017.

Young, T., D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75, 2018.

Zeiler, M. D. ADADELTA: An Adaptive Learning Rate Method. *CoRR*, abs/1212.5701, 2012. URL http://arxiv.org/abs/1212.5701.

Zwicklbauer, S., C. Seifert, and M. Granitzer. DoSeR - A Knowledge-Base-Agnostic Framework for Entity Disambiguation Using Semantic Embeddings. In *The Semantic Web. Latest Advances and New Domains: 13th International Conference, ESWC 2016,*

Heraklion, Crete, Greece, May 29 – June 2, 2016, Proceedings, pages 182–198, Cham, 2016. Springer International Publishing. ISBN 978-3-319-34129-3.