Semantic Web for Machine Translation: Challenges and Directions

Diego Moussallem[1], Matthias Wauer[1], and Axel-Cyrille Ngonga Ngomo[1]

Data Science Group, University of Paderborn, Germany
first.lastname@upb.de

Abstract. A large number of machine translation approaches have recently been developed to facilitate the fluid migration of content across languages. However, the literature suggests that many obstacles must still be dealt with to achieve better automatic translations. One of these obstacles is lexical and syntactic ambiguity. A promising way of overcoming this problem is using Semantic Web technologies. This article is an extended abstract of our systematic review on machine translation approaches that rely on Semantic Web technologies for improving the translation of texts. Overall, we present the challenges and opportunities in the use of Semantic Web technologies in Machine Translation. Moreover, our research suggests that while Semantic Web technologies can enhance the quality of machine translation outputs for various problems, the combination of both is still in its infancy.

Keywords: Machine Translation · Semantic Web · Knowledge Graphs.

1 Introduction

Alongside increasing globalization comes a greater need for readers to understand texts in languages foreign to them. For example, approximately 48% of the pages on the Web are not available in English[1]. The technological progress of recent decades has made both the distribution and access to content in different languages ever simpler. Translation aims to support users who need to access content in a language in which they are not fluent [9].

However, translation is a difficult task due to the complexity of natural languages and their structure [9]. 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 [3]. From a formal point of view, this means that the goal of MT is to transfer the semantics of text from an input language to an output language [7].

Although MT systems are now popular on the Web, they still generate a large number of incorrect translations. Recently, Popović [19] has classified five types of errors that still remain in MT systems. According to research, the two main faults

[1] https://www.internetworldstats.com/stats7.htm

that are responsible for 40% and 30% of problems respectively, are reordering errors and lexical and syntactic ambiguity. Thus, addressing these barriers is a key challenge for modern translation systems. A large number of MT approaches have been developed over the years that could potentially serve as a remedy. For instance, translators began by using methodologies based on linguistics which led to the family of Rule-Based Machine Translation (RBMT). 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.

Statistical Machine Translation (SMT) and Example-Based Machine Translation (EBMT) were developed to deal with the scalability issue in RBMT [4], a necessary characteristic of MT systems that must deal with data at Web scale. Presently, these approaches have begun to address the drawbacks of rule-based approaches. However, some problems that had already been solved for linguistics based methods reappeared. The majority of these problems are connected to the issue of ambiguity, including syntactic and semantic variations [9]. Nowadays, a novel SMT paradigm has arisen called Neural Machine Translation (NMT) which relies on Neural Network (NN) algorithms. NMT has been achieving impressive results and is now the state-of-the-art in MT approaches. However, NMT is still a statistical approach sharing some semantic drawbacks from other well-defined SMT approaches[10].

One possible solution to address the remaining issues of MT lies in the use of Semantic Web Technologies (SWT), which have emerged over recent decades as a paradigm to make the semantics of content explicit so that it can be used by machines. It is believed that explicit semantic knowledge made available through these technologies can empower MT systems to supply translations with significantly better quality while remaining scalable. In particular, the disambiguated knowledge about real-world entities, their properties and their relationships made available on the Linked Data (LD) Web can potentially be used to infer the right meaning of ambiguous sentences or words.

According to our survey [16], the obvious opportunity of using SWT for MT has already been studied by a number of approaches, especially w.r.t. the issue of ambiguity. In this paper, we present the challenges and opportunities in the use of SWT in MT for translating texts.

2 Related Works

The idea of using a structured Knowledge Base (KB) in MT systems started in the 90s with the work of Knight and Luk [8]. Still, only a few researchers have designed different strategies for benefiting of structured knowledge in MT architectures [2]. Recently, the idea of using Knowledge Graph (KG) into MT systems has gained renewed attention. Du et al. [6] created an approach to address the problem of out-of-vocabulary (OOV) words by using BabelNet [18]. Their approach applies different methods of using BabelNet. In summary, they create additional training data and apply a post-editing technique, which replaces the

OOV words while querying BabelNet. Shi et al. [21] have recently built a semantic embedding model reliant upon a specific KB to be used in NMT systems. The model relies on semantic embeddings to encode the key information contained in words to translate the meaning of sentences correctly. The work consists of mapping a source sentence to triples, which are then used to extract the intrinsic meaning of words to generate a target sentence. This mapping results in a semantic embedding model containing KB triples, which are responsible for gathering the key information of each word in the sentences.

3 Open MT Challenges

The most problematic unresolved MT challenges, from our point of view, which are still experienced by the aforementioned MT approaches are the following:

1. *Complex semantic ambiguity*: This challenge is mostly caused by the existence of homonymous and polysemous words. Given that a significant amount of parallel data is necessary to translate such words and expressions adequately. MT systems commonly struggle to translate these words correctly, even if the models are built upon from 5- or 7-grams. For example, "John promises to keep his room tidy" and "John has some promises to keep until he is trusted again". Although the meaning of both are clear to humans, these sentences for SMT systems are statistically expensive and prone to failure[2]. Additionally, for translating the simple word "bank", context information is essential for determining which meaning to assign to it.
2. *Structural divergence*: By definition, structural reordering is reorganizing the order of the syntactic constituents of a language according to its original structure.It in turn becomes a critical issue because it is the core of the translation process. Every language has its own syntax, thus each MT system needs to have adequate models for the syntax of each language. 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 and also, one English word often groups multiple meanings of Japanese characters. For example, Japanese characters make subtle distinctions between homonyms that would not be clear in a phonetic language such as English.
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

[2] See a complete discussion about the problem: `http://tinyurl.com/yck5ngj8`

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, there are five MT open challenges posed by Lopez and Post [11] which we describe more generically below.

(1) Excessive focus on English and European languages as one of the involved languages in MT approaches and poor research on low-resource language pairs such as African and/or South American languages. (2) The limitations of SMT approaches for translating across domains. Most MT systems exhibit good performance on law and the legislative domains due to the 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) How to translate the huge amount of data from social networks that uniquely deal with no-standard speech texts from users (e.g., tweets). (4) The difficult translations among morphologically rich languages. This challenge shares the same problem with the first one, namely that most research work focuses on English as one of the involved languages. Therefore, MT systems which translate content between, for instance, Arabic and Spanish are rare. (5) For the speech translation task, the parallel data for training 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. Since NMT has shown impressive results on reordering, the main problem turns out to be the disambiguation process (both syntactically and semantically) in SMT approaches [9].

4 Suggestions and Possible Directions using SW

Based on the surveyed works on our research [16], SWT have mostly been applied at the semantic analysis step, rather than at the other stages of the translation process, due to their ability to deal with concepts behind the words and provide knowledge about them. As SWT have developed, they have increasingly been able to resolve some of the open challenges of MT. They may be applied in different ways according to each MT approach.

Disambiguation. Human language is very ambiguous. Most words have multiple interpretations depending on the context in which they are mentioned. In the MT field, Word Sense Disambiguation (WSD) techniques are concerned with 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 [3] stated that an MT system is not able to find the right meaning without a specific knowledge. Although the ambiguity problem has been lessened significantly since the contribution of Carpuat and subsequent works [5], this problem still remains a challenge. As seen in Moussallem et al. [16], MT systems still try to resolve this problem by using domain specific language models to prefer domain specific expressions, but when translating a highly ambiguous sentence or a short text which covers multiple domains, the languages models are not enough.

Semantic Web (SW) has already shown its capability for semantic disambiguation of polysemous and homonymous words. However, SWT were applied in two ways to support the semantic disambiguation in MT. First, the ambiguous words were recognized in the source text before carrying out the translation, applying a pre-editing technique. Second, SWT were applied to the output translation in the target language as a post-editing technique. Although applying one of these techniques has increased the quality of a translation, both techniques are tedious to implement when they have to translate common words instead of named entities, then be applied several times to achieve a successful translation.

The real benefit of SW comes from its capacity to provide unseen knowledge about emergent data, which appears every day. Therefore, we suggest performing the topic-modelling technique over the source text to provide a necessary context before translation. Instead of applying the topic-modeling over the entire text, we would follow the principle of communication (i.e from 3 to 5 sentences for describing an idea and define a context for each piece of text. Thus, at the execution of a translation model in a given SMT, we would focus on every word which may be a homonymous or polysemous word. For every word which has more than one translation, a SPARQL query would be required to find the best combination in the current context. Thus, at the translation phase, the disambiguation algorithm could search for an appropriate word using different SW resources such as DBpedia, in consideration of the context provided by the topic modelling. The goal is to exploit the use of more than one SW resource at once for improving the translation of ambiguous terms. The use of two or more SW resources simultaneously has not yet been investigated.

On the other hand, there is also a syntactic disambiguation problem which as yet 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, implementations of NMT for some non-European languages have not been fully exploited (e.g., Brazilian Portuguese, Latin-America Spanish, Zulu, Hindi) due to the lack of large bilingual data sets on the Web to be trained on. Thus, we suggest gathering relationships among properties within an ontology by using the reasoning technique for handling this issue. For instance, the sentence “Anna usually put her notebook on the table for studying” may be annotated using a certain vocabulary and represented 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 MT system. This sentence usually fails when translated to rich morphological languages, such as Brazilian-Portuguese and Arabic, for which the verb influences the translation of “usually” to the past tense. In this case, a reasoning technique may support the problem of finding a certain rule behind relationships between source and target texts in the alignment phase (training phase). However, a well-known problem of reasoners is the poor run-time performance. Therefore, this run-time deficiency needs to be addressed or minimized before implementing reasoners successfully into MT systems.

Named Entities. Most Named Entity Recognition and Disambiguation (NERD) approaches link recognized entities with database entries or websites. This method helps to categorize and summarize text, but also contributes to the disambiguation of words in texts. The primary issue in MT systems 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. Named Entities are a common and difficult problem in both MT (see Koehn [9]) and SW fields. The SW achieved important advances in NERD using structured data and semantic annotations, e.g., by adding an `rdf:type` statement which identifies whether a certain kiwi is a fruit [15]. In MT systems, however, this problem is directly related to the ambiguity problem and therefore has to be resolved in that wider context.

Although MT systems include good recognition methods, they still need improvement. When an MT system does not recognize an entity, the translation output often has poor quality, immediately deteriorating the target text readability. Therefore, we suggest recognizing such entities before the translation process and first linking them to a reference knowledge base. Afterwards, the type of entities would be agglutinated along with their labels and their translations from a reference knowledge base. For instance, in NMT, the idea is to include in the training set for the aforementioned word "Kiwi", "Kiwi.animal.link, Kiwi.person.link, Kiwi.food.link" then finally to align them with the translations in the target text. For example, in SMT, the additional information can be included by Extensible Markup Language (XML) or by an additional model. In contrast, in NMT, this additional information can be used as parameters in the training phase. This method would also contribute to OOV mistakes regarding names. This idea is supported by [21] where the authors encoded the types of entities along with the words to improve the translation of sentences between Chinese-English. Recently, Moussallem et al. [13] have shown promising results by applying a multilingual entity linking algorithm along with knowledge graph embeddings into the translation phase of a neural machine translation model for improving the translation of entities in texts. Their approach achieved significant and consistent improvements of +3 BLEU, METEOR and CHRF3 on average on the newstest datasets between 2014 and 2018 for WMT English-German translation task.

Non-standard speech. The non-standard language problem is a rather important one in the MT field. Many people use the colloquial form to speak and write to each other on social networks. Thus, when MT systems are applied on this context, the input text frequently contains slang, Multiword Expressions (MWE), and unreasonable abbreviations such as “Idr = I don’t remember.” and “cya = see you”. Additionally, idioms contribute to this problem, decreasing the translation quality. Idioms often have an entirely different meaning than their separated word meanings. Consequently, most translation outputs of such expressions contain errors. For a good translation, the MT system needs to recognize such slang and try to map it to the target language. Some SMT systems like Google or Bing have recognition patterns over non-standard speech from old translations through the Web using SMT approaches. In rare cases SMT can solve this problem, but considering that new idiomatic expressions appear every day and most of them are isolated sentences, this challenge still remains open. Moreover, each person has their own speaking form.

Therefore, we suggest that user characteristics can be applied as context for solving the non-standard language problem. These characteristics can be extracted from social media or user logs and stored as user properties using SWT, e.g., FOAF vocabulary. These ontologies have properties which would help identify the birth place or the interests of a given user. For instance, the properties *foaf:interest* and *sioc:topic* can be used to describe a given person’s topics of interest. If the person is a computer scientist and the model contains topics such as “Information Technology” and “Sports”, the SPARQL queries would search for terms inserted in this context which are ambiguous. Furthermore, the property *foaf:based_near* may support the problem of idioms. Assuming that a user is located in a certain part of Russia and he is reading an English web page which contains some idioms, this property may be used to gather appropriate translations of idioms from English to Russian using a given RDF KB. Therefore, an MT system can be adapted to a user by using specific data about him in Resource Description Framework (RDF) along with given KBs. Recently, Moussallem et al [14] have released a multilingual linked idioms dataset as a first part of supporting the investigation of this suggestion. The dataset contains idioms in 5 languages and are represented by knowledge graphs which facilitates the retrieval and inference of translations among the idioms.

Translating KBs. According to our research, it is clear that SWT may be used for translating KBs in order to be applied in MT systems. For instance, some content provided by the German Wikipedia version are not contained in the Portuguese one. Therefore, the semantic structure (i.e., triples) provided by DBpedia versions of these respective Wikipedia versions would be able to help translate from German to Portuguese. For example, the terms contained in triples would be translated to a given target language using a dictionary containing domain words. This dictionary may be acquired in two different ways. First, by performing localisation, as in the work by J. P. McCrae [12] which translates the terms contained in a monolingual ontology, thus generating a bilingual ontology. Second, by creating embeddings of both DBpedia versions in order to determine the similarity between entities through their vectors. This insight is supported by some recent works, such as Ristoski et al. [20], which creates bilingual embeddings using RDF based on Word2vec algorithms. Therefore, we suggest investigating an MT approach mainly based on SWT using NN for translating KBs. Once the KBs are translated, we suggest including them in the language models for improving the translation of entities.

Besides C. Shi et al [21], Arčan and Buitelaar [1] presented an approach to translate domain-specific expressions represented by English KBs in order to make the knowledge accessible for other languages. They claimed that KBs are mostly in English, therefore they cannot contribute to the problem of MT to other languages. Thus, they translated two KBs belonging to medical and financial domains, along with the English Wikipedia, to German. Once translated, the KBs were used as external resources in the translation of German-English. The results were quite appealing and the further research into this area should be undertaken. Recently, Moussallem et al [17] created THOTH, an approach which translates and enriches knowledge graphs across languages. Their approach relies on two different recurrent neural network models along with knowledge graph embeddings. The authors applied their approach on the German DBpedia with the German translation of the English DBpedia on two tasks: fact checking and entity linking. THOTH showed promising results with a translation accuracy of 88.56 while being capable of improving two NLP tasks with its enriched-German KG .

5 conclusion

In this extended abstract, we detailed the results of a systematic literature review of MT using SWT for improving the translation of natural language sentences. Our goal was to present the current open MT translation problems and how SWT can address these problems and enhance MT quality. Considering the decision power of SWT, they cannot be ignored by future MT systems. As a next step, we intend to continue elaborating a novel MT approach which is capable of simultaneously gathering knowledge from different SW resources and consequently being able to address the ambiguity of named entities and also contribute to the OOV words problem. This insight relies on our recent works, such as [13], which have augmented NMT models with the usage of external knowledge for improving the translation of entities in texts. Additionally, future works that can be expected from fellow researchers, include the creation of multilingual linguistic ontologies describing the syntax of rich morphologically languages for supporting MT approaches. Also, the creation of more RDF multilingual dictionaries which can improve some MT steps, such as alignment.

Acknowledgments

This work was supported by the German Federal Ministry of Transport and Digital Infrastructure (BMVI) in the projects LIMBO (no. 19F2029I) and OPAL (no. 19F2028A) as well as by the Brazilian National Council for Scientific and Technological Development (CNPq) (no. 206971/2014-1)

References

1. Arcan, M., Buitelaar, P.: Translating domain-specific expressions in knowledge bases with neural machine translation. CoRR, abs/1709.02184 (2017)
2. Arcan, M., Turchi, M., Buitelaar, P.: Knowledge Portability with Semantic Expansion of Ontology Labels. In: ACL. pp. 708–718 (2015)
3. Bar-Hillel, Y.: The present status of automatic translation of languages. In: Advances in computers, vol. 1, pp. 91–163. Elsevier (1960)
4. Brown, P.F., Cocke, J., Della Pietra, S.A., Della Pietra, V.J., Jelinek, F., Lafferty, J.D., Mercer, R.L., Roossin, P.S.: A statistical approach to machine translation. Computational linguistics **16**(2), 79–85 (1990)
5. Carpuat, M., Wu, D.: Improving Statistical Machine Translation Using Word Sense Disambiguation. In: EMNLP-CoNLL. vol. 7, pp. 61–72 (2007)
6. Du, J., Way, A., Zydron, A.: Using babelnet to improve OOV coverage in SMT. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation 2016. pp. 9–15 (2016)
7. Hutchins, W.J., Somers, H.L.: An introduction to machine translation, vol. 362. Academic Press London (1992)
8. Knight, K., Luk, S.K.: Building a large-scale knowledge base for machine translation. In: AAAI. vol. 94, pp. 773–778 (1994)
9. Koehn, P.: Statistical Machine Translation. Cambridge University Press (2010)
10. Koehn, P., Knowles, R.: Six Challenges for Neural Machine Translation. arXiv preprint arXiv:1706.03872 (2017)
11. Lopez, A., Post, M.: Beyond bitext: Five open problems in machine translation. In: Proceedings of the EMNLP Workshop on Twenty Years of Bitext. pp. 1–3 (2013)
12. McCrae, J.P., Arcan, M., Asooja, K., Gracia, J., Buitelaar, P., Cimiano, P.: Domain adaptation for ontology localization. Web Semantics: Science, Services and Agents on the WWW pp. 23–31 (2016)
13. Moussallem, D., Arčan, M., Ngomo, A.C.N., Buitelaar, P.: Augmenting neural machine translation with knowledge graphs. arXiv preprint arXiv:1902.08816 (2019)
14. Moussallem, D., Sherif, M.A., Esteves, D., Zampieri, M., Ngomo, A.C.N.: LIdioms: A Multilingual Linked Idioms Data Set. In: LREC 2018. p. 7 (2018)
15. Moussallem, D., Usbeck, R., Röeder, M., Ngomo, A.C.N.: MAG: A Multilingual, Knowledge-base Agnostic and Deterministic Entity Linking Approach. In: Proceedings of the Knowledge Capture Conference. p. 9. ACM (2017)
16. Moussallem, D., Wauer, M., Ngomo, A.C.N.: Machine translation using semantic web technologies: A survey. Journal of Web Semantics **51**, 1–19 (2018)
17. Moussallem, D., et al.: THOTH: Neural Translation and Enrichment of Knowledge Graphs. In: The Semantic Web ISWC 2019, pp. 1–17. Springer (2019)
18. Navigli, R., Ponzetto, S.P.: BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. Artificial Intelligence **193**, 217–250 (2012)
19. Popović, M.: Class error rates for evaluation of machine translation output. In: Proceedings of the Seventh Workshop on Statistical Machine Translation. pp. 71–75. Association for Computational Linguistics (2012)
20. Ristoski, P., Paulheim, H.: Rdf2vec: Rdf graph embeddings for data mining. In: International Semantic Web Conference. pp. 498–514. Springer (2016)
21. Shi, C., et al.: Knowledge-Based Semantic Embedding for Machine Translation. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics). vol. 1, pp. 2245–2254 (2016)