Procesamiento del Lenguaje Natural, Revista nº 66, marzo de 2021, pp. 197-200 recibido 13-12-2020 revisado 24-01-2021 aceptado 03-02-2021

Document-Level Machine Translation – Ensuring Translational Consistency of Non-Local Phenomena

Traducción Automática a Nivel de Documento – Asegurando la Consistencia en la Traducción de Fenómenos no Locales

Eva Martínez Garcia
TALP Research Center
Univesitat Politècnica de Catalunya
Jordi Girona, 1-3, 08034 Barcelona, Spain
martinezgarcia.eva@gmail.com

Abstract: PhD Thesis written by Eva Martínez Garcia under the supervision of Dr. Cristina España-Bonet and Dr. Lluís Màrquez. The thesis was defended at the Universitat Politècnica de Catalunya in Barcelona on the 19th of December, 2019. The doctoral committee comprised of Dr. Kepa Sarasola (President, University of Basque Country (UPV/EHU)), Dr. Marta Ruiz Costa-Jussà (Universitat Politècnica de Catalunya (UPC)) and Dr. Sara Stymne (Uppsala Universitet). The thesis was awarded an excellent grade and international mention.
Keywords: Machine Translation, Document-level, Context-aware translation.

Resumen: Tesis doctoral elaborada por Eva Martínez Garcia bajo la supervisión de los doctores Cristina España-Bonet y Lluís Màrquez. La defensa de la tesis tuvo lugar en la Universitat Politècnica de Catalunya en Barcelona el 19 de diciembre de 2019. El tribunal estuvo compuesto por los doctores Kepa Sarasola (Presidente, Universidad del País Vasco (UPV/EHU)), Marta Ruiz Costa-Jussà (Universitat Politècnica de Catalunya (UPC)) y Sara Stymne (Uppsala Universitet). La tesis obtuvo la calificación de sobresaliente y la mención internacional.
Palabras clave: Traducción Automática, Nivel de documento, traducción co contexto.

1 Introduction

Machine Translation (MT) is very present in our daily lives. We use it to access information in other languages on the Internet or to figure out how to say something in languages we do not master for interaction and communication purposes. We are frequent users of the most popular online translation services (e.g., Google Translate, Bing, or DeepL) and we are also used to consuming the MT services provided by social networks (e.g., Facebook or Twitter), which allow us to access the published information in our preferred language. MT is present even in telecommunication applications like Skype, which offers video chats with real-time speech-to-speech translation services. This extended use of MT technology makes us familiarized with its advantages and drawbacks.

Although current MT systems have achieved good translation quality, even comparable with human translation quality in some cases (Wu et al., 2016; Hassan et al., 2018), they still hold a known limitation: they work at sentence level. MT systems translate a document sentence by sentence, taking into account a short context and ignoring document-level information. For all kinds of systems, ignoring extra-sentential information is required due to performance concerns and to the difficulty of properly representing long-distance dependencies. Statistical Machine Translation (SMT) systems (Koehn et al., 2007) rely on local n-gram information, and for Neural Machine Translation (NMT) systems (Bahdanau, Cho, and Bengio, 2015; Vaswani et al., 2017) it is still an open problem how to represent long sequences of words with fixed-length vectors. Thus, state-of-the-art systems perform translation assuming that every sentence can be translated in an isolated way.

ISSN 1989-7553. DOI 10.26342/2021-66-17

However, texts contain relationships among words that hold their coherence, cohesion, and consistency across sentences (Dijk, 1977; Sanders and Pander Maat, 2006).We consider that a good translation should reflect and maintain these qualities at the same degree as they appear in the source text. This is the motivation for our work, which explores techniques to improve the coherence and cohesion levels of the translations generated by state-of-the-art MT systems. Some of the typical mistakes of current MT systems can be linked to the lack of contextual coherence present in the followed translation approaches. We take as inspiration how human translators can resolve these phenomena naturally, by using the entire document's context information.

As an example to illustrate this phenomenon, consider using an MT system to translate a news item in English about a claim process in some office. The word "desk" can appear several times and it can be translated into Spanish as "mostrador", "ventanilla", "escritorio", or "mesa". These Spanish words are not synonyms. Where "mostrador" and "ventanilla" can both be a counter where a service is offered, "mesa" and "escritorio" refer to a piece of furniture. So, "desk" is a word with ambiguous translation into Spanish. Within the context of our example, "mesa" and "escritorio" are not correct translations for "desk". We address this as a contextual coherence problem. Our work aim is to use inter-sentence context to help the system choose a more adequate translation without any knowledge from the domain.

Another typical issue is word agreement across translation segments. Coreference chains confer cohesion to a document, and it is desirable to see this property projected into the produced translations. Unfortunately, this is a property that is typically difficult to maintain for MT systems. Also, gender and number agreement between words is sometimes challenging for current MT approaches. For example, consider the following set of sentences in a source document in English: "She studied civil engineering. [...] The civil engineer was the youngest in the company." These sentences can be translated into Spanish as "*Ella* estudió ingeniería civil. [...] *El ingeniero* era *el* más joven de la empresa." This translation is correct in Spanish if we look at it sentence by sentence.However, it is incorrect if we consider it in its entirety as part of the same document, since there is no gender agreement between the translations of "the engineer" and "she". Taking document context into account, the correct translation would be "*Ella* estudió ingeniería civil. [...] *La ingeniera* era *la* más joven de la empresa."

Our work is motivated by the idea that exploiting discourse information would help to improve the quality of the resulting machine translations at document level. All the techniques we explore in this thesis attempt to find the best way to exploit such kind of information within the current MT frameworks.

2 Research Goals

The general goal of the thesis' work is to improve MT quality by exploring the use of document-level information at different steps of the translation process in order to fix or prevent some of the errors made by sentence-level MT systems. The main goal is to improve machine translation coherence and cohesion by leveraging the information given by the relations of the words along a document.

We define a research strategy with the following steps:

1. *Analyzing translation errors related to document-level phenomena and designing simple methods to tackle them.* A first step towards improving document-level machine translation is to identify those phenomena that confer coherence and cohesion to documents and are susceptible to be lost in the MT process. Before solving such mistakes during the MT process, it is interesting to implement a set of simple post-processing techniques and evaluate their impact.
2. *Capturing the semantic information of a document in a useful manner to aid the MT decoding process.* Leveraging a document's semantic context should help improve the coherence and cohesion levels of its translation. It is necessary to explore ways to introduce contextual semantic information into the MT process. Our final intention is to *extend a document-oriented decoder to incorporate document context semantics.*
3. *Enhancing an NMT framework using context-aware techniques.* To finalize, one of our goals is to integrate the explored ideas into the NMT paradigm.

3 Thesis Overview

The thesis is organized in 7 chapters, followed by an appendix.

Chapter 1 introduces and motivates the work, highlighting the importance of using the context information to improve translationś quality. Chapter 2 revisits the state-of-the-art of the MT research area, focusing on the main technologies of the SMT and NMT paradigms, both at sentence and document level. Chapters 3 to 6 present our results.

In Chapter 3, we analyze some of the translation errors related to document-level phenomena and present a set of post-processing strategies to handle them.

Chapter 4 describes how to use word embeddings for decoding. First, we study the applicability of word embeddings to enhance the MT process. Then, we explain a method to enhance a document-oriented SMT decoder with word embeddings working as Semantic Space Language Models.

Next, Chapter 5 describes our extension of a document-oriented SMT decoder to handle the phenomenon of lexical choice consistency. We present a new feature function that guides the decoder towards more lexically consistent translation candidates, as well as a new strategy to shortcut the exploration of the search space.

Chapter 6 presents our approach to extend the NMT decoding process to take into account contextual semantics. In particular, we extend the beam search decoding algorithm by fusing the discourse information captured by the models described in Chapter 4 to work in tandem with the NMT model.

Finally, Chapter 7 draws the conclusions and describes possible avenues of future work.

Additionally, Appendix *A* describes a new document-level decoding strategy based on a swarm optimization algorithm, integrated into the decoder used in Chapter 4 and 5 as an alternative to its default hill climbing.

4 Main Contributions

The set of contributions is as follows:

- *Analysis of translation errors related to document-level phenomena and the development of a set of simple, yet effective, post-processing techniques to handle them.* Since the particular document-level phenomena they handle are sparse, we need a manual evaluation to assess their effectiveness because the automatic evaluation metrics do not capture their improvements. These findings were published as a technical report (Martínez Garcia, España-Bonet, and Màrquez, 2014b) and presented in the SEPLN2014 conference (Martínez Garcia, España-Bonet, and Màrquez, 2014a).

- *Demonstrating that bilingual word embeddings are capable of modeling semantic relations that help the SMT process.* We observe that the quality of the translation and alignments previous to building the semantic models are crucial for the final performance of the embeddings. Word embeddings prove to be helpful in the task of lexical substitution for words that are ambiguously translated within a document. This work resulted in a publication in the SSST-8 conference (Martínez Garcia et al., 2014).

- *Showing that the introduction of bilingual word embeddings guides document-oriented SMT decoders towards more coherent and cohesive translations.* Although we only observe a slight improvement in the results of automatic evaluation metrics, the improvement is consistent among metrics and is larger as we introduce more semantic information, getting the best results when using the models with bilingual information. This approach was presented in the EAMT2015 conference (Martínez Garcia, España-Bonet, and Màrquez, 2015).

- *Designing new strategies that guide document-oriented decoders through the translation search space towards more consistent, coherent, and cohesive translations, focusing on maintaining lexical consistency.* Our strategies based on word embeddings aid the decoder to assess the compatibility of the possible translations for ambiguous words with their context. This extension led to participating in the EAMT2017 conference (Martínez Garcia et al., 2017).

- *Enhancing the NMT decoding algorithm to include contextual semantics captured by a language model based on word embeddings.* We show how the semantic language models can help NMT systems to produce better translations. Our ap-

proach does not need to modify the training process, so we do not need to increase the training time or document-level annotated data. This work was presented in the DiscoMT2019 (Martínez Garcia, Creus, and España-Bonet, 2019).

Acknowledgements

The thesis work was partially supported by an FPI 2010 grant from the Spanish Ministry of Science and Innovation (MICINN) within the OpenMT-2 project (ref. TIN2009-14675-C03-03) of MICINN, a mobility EEBB 2013 grant from the Spanish Ministry of Economy and Competitiveness (MINECO) for a stay at the Department of Linguistics and Philology at the Uppsala University, and by the TACARDI project (ref. TIN2012-38523-C02-02) of the MINECO.

References

Bahdanau, D., K. Cho, and Y. Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations*.

Dijk, T. A. v. 1977. *Text and context: Explorations in the semantics and pragmatics of discourse*. Number 21 in Longman Linguistics Library. Longman.

Hassan, H., A. Aue, C. Chen, V. Chowdhary, J. Clark, C. Federmann, X. Huang, M. Junczys-Dowmunt, W. Lewis, M. Li, S. Liu, T. Liu, R. Luo, A. Menezes, T. Qin, F. Seide, X. Tan, F. Tian, L. Wu, S. Wu, Y. Xia, D. Zhang, Z. Zhang, and M. Zhou. 2018. Achieving human parity on automatic Chinese to English news translation. *CoRR*, abs/1803.05567.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics on Interactive Poster and Demonstration Sessions*.

Martínez Garcia, E., C. Creus, C. España-Bonet, and L. Màrquez. 2017. Using word embeddings to enforce document-level lexical consistency in machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108.

Martínez Garcia, E., C. Creus, and C. España-Bonet. 2019. Context-aware neural machine translation decoding. In *Proceedings of the 4th Workshop on Discourse in Machine Translation*.

Martínez Garcia, E., C. España-Bonet, and L. Màrquez. 2014a. Document-level machine translation as a re-translation process. *Procesamiento del Lenguaje Natural*, 53.

Martínez Garcia, E., C. España-Bonet, and L. Màrquez. 2014b. Experiments on document level machine translation. Technical Report LSI-14-11-R, Universitat Politècnica de Catalunya, Spain.

Martínez Garcia, E., C. España-Bonet, and L. Màrquez. 2015. Document-level machine translation with word vector models. In *Proceedings of the 18th Annual Conference of the European Association for Machine Translation (EAMT)*.

Martínez Garcia, E., C. España-Bonet, J. Tiedemann, and L. Màrquez. 2014. Word's vector representations meet machine translation. In *Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation*.

Sanders, T. J. M. and H. L. W. Pander Maat. 2006. Cohesion and coherence: Linguistic approaches. In *Encyclopedia of Language & Linguistics*. Elsevier.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st Annual Conference on Neural Information Processing Systems*.

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Ł. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.