On the Evaluation of Terminology Translation Errors in NMT and PB-SMT In the Legal Domain: A Study on the Translation of Arabic Legal Documents into English and French

Khadija Ait ElFqih[1] and Johanna Monti[1]

[1]UNIOR NLP Research Group, University of Naples 'L'Orientale'
{kaitelfqih, jmonti@unior.it}

Abstract

In the translation process, terminological resources are used to solve translation problems, so information on terminological equivalence is crucial to make the most appropriate choices in terms of translation equivalence. In the context of Machine translation, indeed, neural models have improved the state-of-the-art in Machine Translation considerably in recent years. However, they still underperform in domain-specific fields and in under-resourced languages. This is particularly evident in translating legal terminology for Arabic, where current Machine Translation outputs do not adhere to the contextual, linguistic, cultural, and terminological constraints posed by translating legal terms in Arabic. In this paper, we conduct a comparative qualitative evaluation and comprehensive error analysis on legal terminology translation in Phrase-Based Statistical Machine Translation and Neural Machine Translation in two language pairs: Arabic-English, Arabic-French. We propose an error typology taking the legal terminology translation from Arabic into account. We demonstrate our findings highlighting the strengths and weaknesses of both approaches in the area of legal terminology translation for Arabic. We also introduce a multilingual gold standard dataset that we developed using our Arabic legal corpus. This dataset serves as a reliable benchmark and/or reference during the evaluation process to decide the degree of adequacy and fluency of the Phrase-Based Statistical Machine Translation and Neural Machine Translation systems.

1 Introduction

Machine Translation (MT) is a subfield of computational linguistics that draws its fundamentals from linguistics, computer science, information theory, artificial intelligence, and statistics (Sepesy Maučec & Donaj, 2019). Phrase-Based Statistical Machine Translation (PB-SMT) (Koehn et al., 2003), a predictive modelling approach to MT, was the main paradigm in MT research for more than two decades. Neural Machine Translation (NMT) (Kalchbrenner et al., 2014; Cho et al., 2014; Nishimura & Akiba, 2017; Vaswani et al., 2017), the current paradigm for MT research, is an approach to automatic translation in which a large neural network is trained by deep learning techniques. Over the last five years, there has been incremental progress in the field of NMT (Koehn, 2020; Herold et al., 2022; Almahasees, 2021; Rossi & Carre, 2022) to the point where some researchers claim parity with human translation (Thierry, 2022). Consistent term translation is an important facet of quality assurance for specialized translation. Since terminologies are essential for communication among domain experts, term forms must be consistent and their translation must respond to the contextual requirements to maintain the integrity of the underlying conceptual system during knowledge exchange (Darwish, 2009; Sager, 1990). Nevertheless, some knowledge domains and languages still suffer from the lack of high-quality MT results due to the mistranslation of terminology (Mediouni, 2016; Killman, 2014; Zakraoui et al., 2021). This is the case, especially in the legal domain and the Arabic language. Consider example 1 from the Moroccan family code, taking the terms 'الفراش', 'الطعن', 'اللعان', 'القطع' into consideration:

Proceedings of the First ConTenNTS Workshop and the 16th BUCC workshop, pages 26–35
Varna, Sep 7, 2023
https://doi.org/10.26615/978-954-452-090-8_004

1. **AR:** يعتبر **الفراش** بشروطه حجة قاطعة على ثبوت النسب, لا يمكن **الطعن** فيه إلا من الزوج عن طريق **اللعان**, أو بواسطة خبرة تفيد **القطع**.
   EN (NMT): The **Mattress**, with its conditions, is considered a definitive proof of Paternity, and it can only be **challenged** by the husband through **li'an**, or by means of experience that proves the **severance**.
   EN (HT): The **Marriage consummation** is considered a strong proof of paternity; it can be **rebutted** only by the husband through **accusation** or **certain** evidence.
   FR (PB-SMT): La **literie** selon ses termes est un argument concluant pour établir la filiation, qui ne peut être **contestée** par le mari que par la **baise**, ou par l'expérience de la coupe, par deux conditions : le mari en question apporte une preuve solide de sa demande; Un mandat a été émis pour cette expertise.
   FR (HT): La **consommation du mariage** est considérée comme une preuve solide signifiant la paternité, il ne peut être **réfutée** que par le mari soit à travers **l'accusation** ou bien une **certaine** preuve.

The bold terms in example 1 are domain-specific and context-dependent, so their correct translation requires the consideration of the context, as well as of the cultural, lexical, morphological, and semantic properties of the terms in addition to their equivalences across languages and legal systems (i.e., English, and French), as the HT does. Both NMT and PB-SMT, instead, produce wrong results. This example highlights the main weaknesses of MT, namely lack of terminology resources related to the legal domain for Arabic, the lack of training on Arabic legal texts to render the appropriate equivalences, and the terminology linguistic characteristics of this type of discourse.

In this work, we aim to compare PB-SMT and NMT with reference to terminology translation by carrying out an extensively detailed manual evaluation. We propose an error typology taking the legal terminology translation from Arabic into account. While automatic metrics provide a quick and cost-effective way to evaluate MT output (Zakraoui et al., 2021; Sepesy Maučec & Donaj, 2020), it is not recommended for evaluating terminology translation errors (Izwaini, 2006; Gamal et al., 2022; Haque et al., 2020; Killman, 2014) because they have limitations in assessing the accuracy, quality, legal context, and cultural nuances of legal translations for Arabic. For this reason, we create a multilingual gold standard dataset (AR-EN / AR-FR) using a corpus of judicial documents (i.e., contracts, provisions, codes, decrees) of different Arab countries (Morocco, Algeria, Tunisia, United Arab Emirates, Saudi Arabia, Egypt) specifically created for this experiment. This multilingual dataset is used as a benchmark for evaluating both the NMT and PB-SMT results concerning out-of-context and in-context legal terms. To ensure the quality and reliability of the reference translations of the gold standard dataset, we collaborate with a legal expert and an Arab linguist who are proficient in both the source and target languages.

2 Related Work

Since the introduction of NMT to the MT community, researchers have been analyzing the pros and cons of NMT compared to PB-SMT. Koehn & Knowles (2017) examine several challenges to NMT and give empirical results on how well the technology holds up compared to PB-SMT. To do this, they train both NMT and PB-SMT for German-English on domains that are quite distant from each other (i.e., law (Acquis), Medical (EMEA), IT, Koran (Tanzil), subtitles) obtained from OPUS (Tiedemann, 2012). They note that the output of the NMT system is often quite fluent but completely unrelated to the input, while the PB-SMT output betrays its difficulties with coping with the out-of-domain input by leaving some words untranslated. They conclude that despite the recent successes, NMT must still overcome various challenges, most notably performance in out-of-domain and under-resourced conditions. Zakraoui et al. (2021) conduct a survey related to Arabic MT challenges which they split into two categories, namely linguistic (i.e., morphology richness, syntactic word reordering, Word Sense Disambiguation, vocalization, dialectal variation, gender bias, etc.) and technical (i.e., low-resource language, domain mismatch, Out-Of-Vocabulary, word alignment, sentence length, among others). Several studies including Alsohybe et al (2017); Hadla et al (2014); Han (2016) prove the ineffectiveness of NMT systems, mainly Google Translate (GT) when producing Arabic-English translations. In the context of domain-specific translation, particularly when dealing with legal texts, the problem

escalates significantly. This is mainly due to domain mismatch (Koehn, 2020), which Wang et al., (2020) tackle using multi-domain NMT.

As long as the MT evaluation is concerned, researchers use different metrics such as Word Error Rate (Sai et al., 2022), METEOR (Lavie & Denkowski, 2009; Banerjee & Lavie, 2005), AL-BLEU (Bouamor et al., 2014) metric which extends BLEU (Papineni et al., 2002) to deal with Arabic rich morphology. Nevertheless, Han, (2016) and another recent study by Lee et al., (2023) try to evaluate several automatic metrics, including the above ones. They prove that no conclusions can be drawn on the superior performance of any specific metric over others. They state that while automatic metrics such as BLEU, capture the average case for how well an MT model translates sentences, they do not give insights into which linguistic aspects MT models struggle with producing fluent output. In this regard, some efforts investigate statistical error analysis of MT for Arabic with native speakers so they can review linguistic aspects of MT errors (El Marouani et al., 2020; Al Mahasees 2020), while others use neural networks to detect errors (Madi & Al-Khalifa, 2020) for Arabic texts or to correct them (Watson et al., 2018). In another study on evaluating terminology translation in MT, Haque et al., (2019) examine why the automatic evaluation techniques fail to distinguish term translation in few cases, and identify the reasons (e.g., reordering, and inflectional issues in term translation). In this regard, they propose the TermEval metric for the automatic evaluation of terminology translation in MT. Nevertheless, the proposed metric supports only the English-Hindi pair because of resources limitation.

We now turn our attention to studies related to terminology translation in MT. Haque et al. (2020) investigate legal domain term translation in PB-SMT and NMT with two morphologically divergent languages, English and Hindi. In their experiment, they adopt a technique that semi-automatically creates a gold standard test set from an English-Hindi judicial domain parallel corpus. The sentences of the gold standard test set are translated with their PB-SMT and NMT systems, and the patterns of the terminology translation errors on a sample set of translations is inspected and classified. A comparative evaluation of PB-SMT and NMT on terminology translation is then carried out. They find that NMT is less prone to errors than PB-SMT as far as terminology translation is concerned (8.3% versus 9.9% and 11.5% versus 12.9% error rates in English-Hindi and Hindi-English translation tasks, respectively; differences in error rates are statistically significant). Their empirical results present divergent outcomes in comparison to those reported in several prior investigations (Vintar, 2018; Dugonik et al. 2023; Khazin et al. 2023). In another scenario, Müller et al. (2019) study the performance of PB-SMT and NMT systems on out-of-domain German-English OPUS data and German-Romansh to define five domains (i.e., medical, IT, koran, law, and subtitles). They find that in unknown domains, PB-SMT and NMT suffer from different problems: PB-SMT systems are mostly adequate but not fluent, while NMT systems are mostly fluent but not adequate. For NMT, they identify hallucinations (translations that are fluent but unrelated to the source) as a key reason for low domain robustness. Several studies, including Al-Shehab (2013); Killman (2014); Junczys-Dowmunt et al. (2016); Baruah & Singh (2023), prove that although NMT systems are known to generalize better than phrase-based systems for out-of-domain data, it is unclear how they perform in purely in-domain setting, especially in the legal domain from Arabic where terminology translation remains questionable and subject to continuous post-edition (Alkatheery, 2023). Given all the serious translation issues that Arabic terminology in the legal domain faces, it remains a poorly explored area in MT research. Hence, extensive research efforts are still needed to enhance and refine these aspects.

3 Experiments Set-up and Methodology

To conduct our study, we semi-automatically create a gold standard dataset [1] from our legal corpus that we created using a variety of legal documents (i.e., codes, contracts, provisions, constitutions, and decrees) of different Arab countries. The resource setup is described in detail in Table 3 and in ElFqih et al. (2023), and, to the best of our knowledge, this is the first formalized resource created specifically for assessing the

[1]Available here: https://github.com/Kaitelfqih/Gold-standard-Terminology-Translation-Evaluation-Data-Set

accuracy and adequacy of MT outputs regarding terminology in the legal domain for Arabic against English and French. This terminology resource consists of:

- 1015 out-of-context legal term translated using NMT system (GT) and PB-SMT system (RC),
- 1015 in-context legal term translated using NMT system (GT) and PB-SMT system (RC),
- Manual annotations of NMT and PB-SMT errors (see section 4),
- 1015 Reference translations for both out-of-context and in-context dataset validated by a legal expert.

To address our research objectives, our methodology unfolds four distinct phases, to investigate key aspects of the study. They are as follow:

- The translation of the out-of-context and in-context terms from Arabic to English, and French using GT and RC,
- The extraction of phrases using NooJ grammars[2] (Silberztein, 2015) containing the terms list understudy,
- The production of the reference translations of the legal terms for Arabic according to online gateways of EU laws, including EUR-Lex[3], IATE[4], Juremy[5],

Our reference translations undergo thorough annotation and validation processes conducted by two skilled annotators:

- The first annotator is a legal expert whose language skills are excellent both in the source and the target languages. He validates the translations after checking their degree of accuracy and adequacy in the target languages.
- The second annotator, a native Arabic speaker with a linguistic background meticulously annotates the Part-of-Speech tags, Geographical Usage (following the ISO 20771:2020 standard for Legal translation Requirements, to indicate where a given term is adapted to express a legal practice).

The above steps are important for the sake of placing equivalence references which ensure an adequate and accurate analysis. The annotators possess a deep understanding of legal concepts and the nuances of the Arabic language. Their combined expertise ensures the accuracy and reliability of the annotations present in the dataset. This dual-annotator approach enhances the quality of the data by reducing the chances of errors and inconsistencies, and it provides a standardized point of reference for evaluating PB-SMT and NMT systems objectively and systematically in the area of legal terminology translation for Arabic.

The second phase of the experiment focuses on manual evaluation carried out by a native Arabic speaker. It consists of a systematic analysis where we classify and annotate the errors (see Section 4) of machine-translated out-of-context and in-context legal terms from AR to EN and FR produced by different MT systems (GT and RC). Figure 1 displays the number of terms and sentences containing errors, along with their corresponding percentages in Table 1 and Table 2.

Table 1: Error Types of Machine-Translated Out-of-Context Legal Terms for Arabic.

	Arabic-English		**Arabic-French**	
Errors	NMT	PB-SMT	NMT	PB-SMT
Ambiguity Errors (AE)	63%	62%	58%	56%
Cultural and Legal Systems Relatedness Errors (CLSRE)				
Register Errors (RE)				
Transliteration Errors (TE)	35%	33%	38%	40%
Gender Bias Errors (GBE)				
None of the Above (Ø)	2%	5%	4%	4%

[2] https://nooj.univ-fcomte.fr/

[3] https://eur-lex.europa.eu/browse/eurovoc.html?locale=en

[4] https://iate.europa.eu/home

[5] https://www.juremy.com/

Table 2 : Error Types of Machine-Translated In-Context Legal Terms for Arabic

	Arabic-English		Arabic-French	
Errors	**NMT**	**PB-SMT**	**NMT**	**PB-SMT**
Reordering Errors (RE)	65.2%	63.8%	63.5%	65.5%
Ambiguity Errors (AE)				
Cultural and Legal Systems Relatedness Errors (CLSRE)				
Register Errors (RE)				
Transliteration Errors (TE)	32.8%	31.2%	31.5%	32.5%
Lexical Repetition Errors (LRE)				
Term Drop Errors (TDR)				
Gender Bias Errors (GBE)				
None of the Above (Ø)	2%	5%	5%	2%

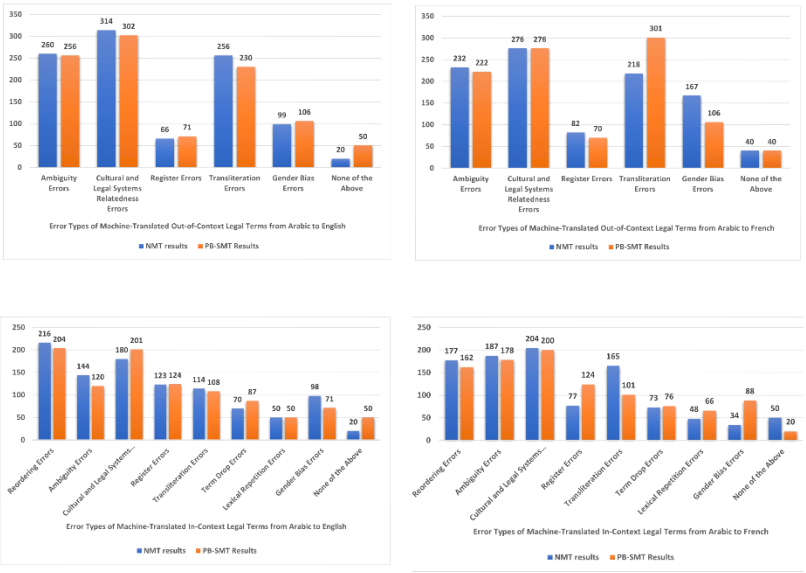

Figure 1: Graphs Showing the Detailed Numbers of Out-of-Context (Upper Graph) and In-Context (Lower Graph) Terms and Their Respective Errors in NMT and PB-SMT Systems from Arabic into English and French

4 Evaluation and Results

The errors posed by machine-translated legal terms for Arabic are classified into six error types for out-of-context terms (Table 1) and in eight types for in-context terms (Table 2).

Table 1 and Table 2 show the manual evaluation results after comparing the outputs of NMT and PB-SMT systems from AR to EN and FR against the gold test-set. Figure 1 provides a detailed overview of the number of both the Out-of-Context and in In-Context legal terms along with the errors identified through the manual evaluation for each context.

Table 1 presents the results of the evaluation of the correspondence between NMT and PB-SMT systems of out-of-context legal terms in AR-EN/AR-FR pairs, indicating the number of terms that contain the errors and their respective percentage.

For AR-EN, NMT appears to be more error-prone than PB-SMT. NMT commits 36% of errors related to AE, CLSRE, RE, and below 40% of errors related to TE and GBE. Whereas PB-SMT presents 62% of errors related to AE, CLSRE, RE, and below 33% of errors related to TE and GBE. In addition, only 2% in NMT and 5% in PB-SMT are correct translations. For the AR-FR pair, NMT preserves its status of being more erroneous than PB-SMT, where NMT presents a percentage of 58% of errors related to AE, CLSRE, RE, and 38% in favor of TE and GBE. Whereas PB-SMT achieves 56% of AE, CLSRE, RE, but outperforms NMT with 40% of errors related to TE and GBE.

In addition, 4% in NMT and 4% in PB-SMT are correct translations.

Table 2 includes the manual evaluation results of in-context machine-translated legal sentences where the terms in Table 1 are spotted. The findings for NMT and PB-SMT for AR-EN/ AR-FR show that the percentage and number of errors obtained after translating the terms in context increase in comparison with the previous results (Table 1) obtained for out-of-context terms. In other words, for AR-EN pairs, NMT seems to exhibit a higher percentage of errors compared to PB-SMT, and vice versa for AR-FR pairs. However, the incidence rate is higher in errors related to RE, AE, CLSRE, RE. The findings reveal that the inclusion of contextual information makes it hard for the MT systems to mitigate these errors and produce accurate legal translations for Arabic, consider example 1:

1. **AR:** **حكم** القاضي بتاريخ 2011/01/01 بجميع ما للزوجة على الزوج من **واجبات** من **نفقة** و **متعة**.
   EN (NMT): The judge **ruled** on 01/01/2011 all the **duties** of the wife to the husband of **maintenance** and **pleasure**.
   EN (HT): On 01/01/2011, the judge **sentenced** that the husband must comply with all the wife's **rights**, including **expenditure** and **compensation**.
   FR (SMT): Le juge a **statué** le 01/01/2011 sur l'ensemble des **devoirs** de la femme envers le mari **d'entretien** et de **jouissance**.
   FR (HT): Le 01/01/2011, le juge a **condamné** le mari à respecter tous les **droits** de sa femme, y compris les **dépenses** et **l'indemnisation**.

The bold terms in example 1 are domain-specific and context-dependent these factors make their accurate translation a complex process. The HT, indeed, considers various elements, including context, exact terminology choice, structure, syntax, as well as their compatibility across languages and legal systems. Whereas NMT and PB-SMT systems fail in producing quality translations due to errors, such as:

- RE, which disrupts the sentence structure, leading to confusion in the intended meaning of legal terms, and which might not align with the conventions of legal writing in Arabic, potentially affecting the legal validity and clarity of the text,
- AE that creates multiple interpretations of legal terms, causing uncertainty and potential misinterpretations in legal documents,
- CLSRE where certain concepts or practices does not exist in the target legal system, leading to inappropriate or misleading translations because legal texts and terms are deeply influenced by the cultural and historical context of the legal system they belong to,
- TDE where MT systems omit the source term in translation,
- LRE when translation of a source term, is an incorrect lexical choice,
- GBE which significantly impacts legal translation from Arabic into English and French, as these languages have different ways of handling gender in their grammatical structures and legal systems.

Our corpus consists of judicial documents (i.e., contracts, provisions, codes, decrees) of different Arab countries (Morocco, Algeria, Tunisia, United Arab Emirates, Saudi Arabia, Egypt). Therefore, the use of distinct legal terminology to convey similar legal practices in different countries can significantly impact the outcomes of MT for Arabic. Due to variations in legal systems, cultural nuances, idiomatic expressions, linguistic variations, and the specific precision required in legal language, MT may struggle to accurately capture the intended meanings. This could lead to mistranslations, misinterpretations, and errors that have potentially serious legal consequences. For example, the term **'مأذون'** is used mostly in Qatar and Egypt. It is used to refer to the person certified by the judge to perform certain legal formalities, especially to draw up or certify marriage contracts, deeds, and other documents for use in other jurisdictions [6] . RC, however, translates it as **'authorized'** into English and **'autorisé'** into French. Whereas GT, as well, translates it as **'authorized'** into English and **'autorisé'** into French. Therefore, we notice that both systems not only transform the grammatical category of the term from a noun, which represents a person into an adjective, but they also misinterpret the intended

[6] https://www.almaany.com/ar/dict/ar-ar/

legal practice in the target legal systems. Hence, in France, the equivalence of **'مأذون'** is **'maire'** (i.e., the person who chairs the municipal council[7]), he/she is the one who oversees approving and drawing up marriage contracts. Whereas in England the person in charge of approving and celebrating the marriage requests is called the **'superintendent registrar[8]'** of the district.
This unveils that MT systems are not trained on a diverse and comprehensive dataset that covers a wide range of legal terminologies from different countries. In other words, MT systems need to be equipped with region-specific legal dictionaries and context-aware algorithms that consider the nuances of each country's legal language. Additionally, leveraging parallel legal texts in different terms can help train MT models to better handle these variations.

5 Conclusion and Future Work

In this paper, we conduct a comparative qualitative evaluation and comprehensive error analysis on legal terminology translation between PB-SMT and NMT in two translation pairs: AR-EN/ AR-FR. We also introduce a multilingual gold standard dataset that we developed using our Arabic legal corpus, which serves as a reliable benchmark and/or reference during the evaluation process to decide the degree of adequacy and fluency of the PB-SMT and NMT systems. We propose an error typology taking the legal terminology translation from Arabic into account.

We demonstrate our findings, highlighting the strengths and weaknesses of both approaches to MT in legal terminology translation for Arabic. We found that NMT is more error-prone than PB-SMT in both language pairs when translating out-of-context terms. Whereas, for the AR-EN pair, NMT seems to exhibit a higher percentage of errors compared to PB-SMT concerning in-context machine-translated legal terms. Concerning the AR-FR language pair, although NMT and PB-SMT have the same overall error rate (94%) NMT produces more errors related to RE, AE, CLSRE, and register errors.

The findings also demonstrate that despite advances in MT, legal translation remains a challenging task that demands precision and adherence to specific legal nuances. For critical legal documents, human translation by professional legal experts is still the preferred approach to ensure the highest level of accuracy and consistency. MT, however, can be a helpful tool for initial draft translations or to aid human translators, but it should be used with caution, especially for legal content.

As future work, a second annotator will undertake the annotation of the data concerning the MT errors, and the assessment of inter-annotator agreement will be conducted to enhance the reliability of the data.
We will afterward focus on developing a high-quality multilingual corpus from AR-EN/ AR-FR in the legal domain to enhance the performance of MT systems. Careful attention will be given to aligning sentences with precise legal terminologies to provide reliable and contextual translations.

References

AlMahasees, Z. (2020). Diachronic evaluation of Google Translate, Microsoft Translator, and Sakhr in English-Arabic translation [Master thesis]. Unpublished Master's Thesis, the University of Western Australia, Australia.

Alkatheery, E. R. (2023). Google translate errors in legal texts: Machine translation quality assessment. Center for Open Science. http://dx.doi.org/10.31235/osf.io/j4zh7

Al-Rukban, A., & Saudagar, A. K. J. (2017, December 20). Evaluation of English to Arabic Machine Translation Systems using BLEU and GTM. Proceedings of the 2017 9th International Conference on Education Technology and Computers. http://dx.doi.org/10.1145/3175536.3175570

Al-Shehab, M. (2013). The translatability of English legal sentences into Arabic by using Google translation. International Journal of English Language and Linguistics Research, 1(3), 18–31.

Alsohybe, N., Dahan, N., & BaAlwi, F. (2017). Machine-Translation history and evolution: Survey for Arabic-english translations. Current Journal of Applied Science and Technology, 23(4), 1–19. https://doi.org/10.9734/cjast/2017/36124

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved

[7]EESC/COR-FR, d'après le Conseil des communes et régions d'Europe (CCRE), «Gouvernements locaux et régionaux en Europe — Structures et compétences» (2016) (3.5.2022), page 26

[8]Term reference: https://www.citizensadvice.org.uk/family/living-together-marriage-and-civil-partnership/getting-married/

correlation with human judgments. Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 65–72.

Baruah, R., & Singh, A. K. (2023). A clinical practice by machine translation on low resource languages. In Natural Language Processing in Healthcare (pp. 1–17). CRC Press. http://dx.doi.org/10.1201/9781003138013-1

Berrichi, S., & Mazroui, A. (2021). Addressing limited vocabulary and long sentences constraints in english–arabic neural machine translation. Arabian Journal for Science and Engineering, 46(9), 8245–8259. https://doi.org/10.1007/s13369-020-05328-2

Bouamor, H., Alshikhabobakr, H., Mohit, B., & Of lazer, K. (2014). A human judgement corpus and a metric for Arabic MT evaluation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). http://dx.doi.org/10.3115/v1/d14-1026

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder–Decoder approaches. Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation. http://dx.doi.org/10.3115/v1/w14-4012

Cuong, H., & Sima'an, K. (2017). A survey of domain adaptation for statistical machine translation. Machine Translation, 31(4), 187–224. https://doi.org/10.1007/s10590-018-9216-8

Darwish, A. (2009). Terminology and translation: A phonological-semantic approach to Arabic terminology. Writescope Publishers.

Dugonik, J., Sepesy Maučec, M., Verber, D., & Brest, J. (2023). Reduction of neural machine translation failures by incorporating statistical machine translation. Mathematics, 11(11), 2484. https://doi.org/10.3390/math11112484

El Marouani, M., Boudaa, T., & Enneya, N. (2020). Statistical error analysis of machine translation: The case of Arabic. Computación y Sistemas, 24(3). https://doi.org/10.13053/cys-24-3-3289

ElFqih, K. A., di Buono, M. P., & Monti, J. Towards a Linguistic Annotation of Arabic Legal Texts: A Multilingual Electronic Dictionary for Arabic. In Book of Abstracts (p. 17).

Gamal, D., Alfonse, M., Jimenez-Zafra, S. M., & Aref, M. (2022, May 8). Survey of Arabic machine translation, methodologies, progress, and challenges. 2022 2nd International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC). http://dx.doi.org/10.1109/miucc55081.2022.9781776

Hadla, L., Taghreed, H., & Al-Kabi, M. (2014). Evaluating Arabic to English Machine Translation. International Journal of Advanced Computer Science and Applications, 5, 68–73. https://doi.org/10.14569/IJACSA.2014.051112#sthash.NW2l6vl5.dpuf

Halimi, S. A. (2017). Contextualizing translation decisions in legal system-bound and international multilingual contexts. Between Specialised Texts and Institutional Contexts – Competence and Choice in Legal Translation, 3(1), 20–46. https://doi.org/10.1075/ttmc.3.1.03hal

Han, L. (2016). Machine translation evaluation resources and methods: A survey. arXiv Preprint arXiv:1605.04515.

Haque, R., Hasanuzzaman, M., & Way, A. (2019). TermEval: An automatic metric for evaluating terminology translation in MT. Springer.

Haque, R., Hasanuzzaman, M., & Way, A. (2020). Analysing terminology translation errors in statistical and neural machine translation. Machine Translation, 34(2–3), 149–195. https://doi.org/10.1007/s10590-020-09251-z

Herold, C., Rosendahl, J., Vanvinckenroye, J., & Ney, H. (2022). Detecting various types of noise for neural machine translation. Findings of the Association for Computational Linguistics: ACL 2022. http://dx.doi.org/10.18653/v1/2022.findings acl.200

Izwaini, S. (2006). Problems of Arabic machine translation: evaluation of three systems. Proceedings of the International Conference on the Challenge of Arabic for NLP/MT, 118–148. https://aclanthology.org/2006.bcs-1.11

Junczys-Dowmunt, M., Dwojak, T., & Hoang, H. (2016). Is neural machine translation ready for deployment? A case study on 30 translation directions. arXiv Preprint arXiv:1610.01108.

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). http://dx.doi.org/10.3115/v1/p14-1062

Khazin, K. M., Sanjaya, D., Siregar, M., Meisuri, & Adisaputra, A. (2023). Comparison of machine translations (MT) technology; statistical (SMT) vs. neural (NMT). ADVANCES IN FRACTURE AND DAMAGE MECHANICS XX. http://dx.doi.org/10.1063/5.0133311

Killman, J. (2014). Vocabulary accuracy of statistical machine translation in the legal context. Proceedings of the 11th Conference of the

Association for Machine Translation in the Americas, 85–98.

Koehn, P. (2020). Neural machine translation. Cambridge University Press.

Koehn, P., & Knowles, R. (2017). Six challenges for neural machine translation. Proceedings of the First Workshop on Neural Machine Translation. http://dx.doi.org/10.18653/v1/w17-3204

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical phrase-based translation. Defense Technical Information Center. http://dx.doi.org/10.21236/ada461156

Lavie, A., & Denkowski, M. J. (2009). The Meteor metric for automatic evaluation of machine translation. Machine Translation, 23(2–3), 105–115. https://doi.org/10.1007/s10590-009-9059-4

Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S., & Lim, H. (2023). A Survey on Evaluation Metrics for Machine Translation. Mathematics, 11(4), 1006.

Madi, N., & Al-Khalifa, H. (2020). Error detection for Arabic text using neural sequence labeling. Applied Sciences, 10 (15), 5279. https://doi.org/10.3390/app10155279

Mediouni, M. (2016). Towards a functional approach to Arabic–english legal translation: The role of comparable/parallel texts. In M. Taibi (Ed.), New Insights into Arabic Translation and Interpreting (pp. 115–160). Multilingual Matters. http://dx.doi.org/10.21832/9781783095254-008

Müller, M., Rios, A., & Rico, S. (2019). Domain robustness in neural machine translation. arXiv Preprint arXiv:1911.03109.

Nishimura, T., & Akiba, T. (2017, August). Addressing unknown word problem for neural machine translation using distributee representations of words as input features. 2017 International Conference on Advanced Informatics, Concepts, Theory, and Applications (ICAICTA). http://dx.doi.org/10.1109/icaicta.2017.8090977

Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02. http://dx.doi.org/10.3115/1073083.1073135

Rossi, C., & Carre, A. (2022). Machine translation for everyone: Empowering users in the age of artificial intelligence. Language Science Press Berlin, 18, 51. https://doi.org/10.5281/zenodo.6653406

Sager, J. C. (1990). Practical course in terminology processing. John Benjamins Publishing.

Sai, A. B., Mohankumar, A. K., & Khapra, M. M. (2022). A survey of evaluation metrics used for NLG systems. ACM Computing Surveys, 55(2), 1–39. https://doi.org/10.1145/3485766

Sepesy Maučec, M., & Donaj, G. (2019). Machine translation and the evaluation of its quality. In A. Sadollah & S. Tilendra (Eds.), Recent Trends in Computational Intelligence. IntechOpen. http://dx.doi.org/10.5772/intechopen.89063

Thierry, P. (2022). On" Human Parity" and" Super Human Performance" in Machine Translation Evaluation. Language Resource and Evaluation Conference.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. Lrec, 2012, 2214–2217.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, June 12). Attention is all you need. arXiv.Org. https://arxiv.org/abs/1706.03762

Vintar, Š. (2018). Terminology translation accuracy in statistical versus neural MT: An evaluation for the English-Slovene language pair. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), 34–37.

Wang, W., Tian, Y., Ngiam, J., Yang, Y., Caswell, I., & Parekh, Z. (2020). Learning a multi-domain curriculum for neural machine translation. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. http://dx.doi.org/10.18653/v1/2020.acl-main.689

Silberztein, M. (2015). La formalisation des langues: l'approche de NooJ. ISTE Group.

Watson, D., Zalmout, N., & Habash, N. (2018). Utilizing character and word embeddings for text normalization with sequence-to-sequence models. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. http://dx.doi.org/10.18653/v1/d18-1097

Zakraoui, J., Saleh, M., Al-Maadeed, S., & AlJa'am, J. M. (2020, April). Evaluation of Arabic to English machine translation systems. 2020 11th International Conference on Information and Communication Systems (ICICS). http://dx.doi.org/10.1109/icics49469.2020.239518

Zakraoui, J., Saleh, M., Al-Maadeed, S., & Alja'am, J. M. (2021). Arabic machine translation: A survey with challenges and future directions. IEEE Access, 9, 161445–161468. https://doi.org/10.1109/access.2021.3132488

Ziemski, M., Junczys-Dowmunt, M., & Pouliquen, B. (2016). The United Nations parallel corpus v1. 0. Proceedings of the Tenth International Conference

on Language Resources and Evaluation (LREC'16),
3530–3534.

Appendix

Table 3: Arabic Legal Documents

Documents	Type	Country	Tokens
Family Code	**Code**	**Morocco**	**20,726**
Code of Penal Procedures	**Code**	**Morocco**	**76,945**
Code of Obligations and Contracts	**Code**	**Morocco**	**82,365**
Civil Code	**Code**	**Algeria**	**113,287**
Penal Code		**Algeria**	**113,287**
Tunisian Code of Penal Status	**Code**	**Tunisia**	**11,638**
Code of Penal Procedures	**Code**	**Tunisia**	**11,638**
Qatari Civil Code	**Code**	**Qatar**	**62,601**
Constitution of the Kingdom of Morocco	**Constitution**	**Morocco**	**12,494**
Marriage Contract	**Contract**	**Morocco**	**315**
Real Estate Sale Contract	**Contract**	**Algeria**	**427**
Divorce by Mutual Consent before Marriage consummation	**Provision**	**Morocco**	**277**
Irrevocable Divorce after Marriage Consummation	**Provision**	**Egypt**	**100**
Irrevocable Divorce before Marriage Consummation	**Provision**	**Egypt**	**131**
Revocable divorce	**Provision**	**Egypt**	**86**
Self-divorce	**Provision**	**Morocco**	**308**
Total of Tokens			**2148,981**