Building a Corpus for Corporate Websites Machine Translation Evaluation

A Step by Step Methodological Approach

Irene Rivera-Trigueros[1][0000-0003-4877-4083] and María-Dolores Olvera-Lobo[2][0000-0002-0489-7674]

[1]University of Granada, Department of Translation and Interpreting, Faculty of Translation and Interpreting, C/ Buensuceso, 11, 18002, Granada, Spain
irenerivera@ugr.es

[2] University of Granada, Department of Information and Communication, Colegio Máximo de Cartuja, Campus Cartuja s/n, 18071, Granada, Spain
molvera@ugr.es

Abstract. The aim of this paper is to describe the process carried out to develop a parallel corpus comprised of texts extracted from the corporate websites of southern Spanish SMEs from the sanitary sector which will serve as the basis for MT quality assessment. The stages for compiling the parallel corpora were: (i) selection of websites with content translated in English and Spanish, (ii) downloading of the HTML files of the selected websites, (iii) files filtering and pairing of English files with their Spanish equivalents, (iv) compilation of individual corpora (EN and ES) for each of the selected websites, (v) merging of the individual corpora into a two general corpus one in English and the other in Spanish, (vi) selection a representative sample of segments to be used as original (ES) and reference translations (EN), (vii) building of the parallel corpus intended for MT evaluation. The parallel corpus generated will serve to future Machine Translation quality assessment. In addition, the monolingual corpora generated during the process could as a base to carry out research focused on linguistic–bilingual or monolingual−analysis.

Keywords: Parallel Corpora, Monolingual Corpora, Machine Translation, Machine Translation Quality Assessment, Corporate Websites.

1 Background

Nowadays, thanks to the development of information and communication technologies, companies are able to spread messages globally, allowing them to open new markets. Web 2.0 tools, such as websites, provides enterprises, especially Small and Medium-sized ones (SMEs) with great opportunities for internationalization [1]. In fact, in the European Union (EU), more than 99% of all enterprises − save for the financial business sector − are SMEs [2] and 77% of them have a website [3]. However, language barriers often pose a challenge for companies when it comes to the multilingual dissemination of corporate information and that is why Machine Transla-

Translation and Interpreting Technology Online, pages 93–101,
July 05-07, 2021.
https://doi.org/10.26615/978-954-452-071-7_011

tion (MT) can be a resource with great potential for solving this problem. Nevertheless, MT quality is generally inferior to that reached by professional human translations. Consequently, MT evaluation, by means of automatic and human metrics, plays a key role for determining MT quality as well as for MT systems to be improved. However, the assessment of MT systems implies cognitive linguistic, social, cultural and technical processes [4]. As a result, assessing MT can present difficulties as in the majority of cases there is not just one correct translation [5]. In addition, it is important to note that there is a great lack of consensus in relation to translation quality assessment and approaches may differ according to the individuals, groups or contexts in which quality is assessed. Therefore, there are a number of metrics and criteria for undertaking the evaluation of MT systems. However, generally speaking, there are two main types of MT evaluation: human and automatic.

On the one hand, most automated metrics establish comparisons between the output of an MT system and one or more reference translation [4, 6]. Some of these measures such as WER, PER or TER are based on the Levenshtein or edit distance [7], the difference among this metrics is that some of this metrics consider word or phrases reordering as and edit operation [8]. Other measures, such as BLEU [9]–the most popular metric–are based on precision and carried out at the level of n-grams, indivisible language units. BLEU employs a modified precision that considers the maximum number of each n-gram appearance in the reference translation and applies a brevity penalty that is added to the measurement calculation. Other precision-centred metrics, for example, are NIST [10], ROUGE [11], F-measure [12] and METEOR [13].

On the other hand, human evaluation revolves around adequacy–based on semantic quality–and fluency–based on syntactic quality. For adequacy, evaluation reference translations or the original text, if the evaluators have language knowledge, are required. In the case of fluency evaluation, as the evaluation is monolingual, no reference translation nor the original text are necessary. Human evaluation can be carried out by means of ranking, Likert-type ordinal scales, gap filling tasks or by identifying, annotating, classifying and correcting translation errors, amongst others [8].

Human evaluation, despite demanding more time, effort and costs, is considered to be more reliable than automatic metrics, as their capacity to evaluate syntactic and semantic equivalence is limited [4, 6]. However, human evaluation cannot be reproduced, is less objective than automatic metrics and requires evaluators to fulfill certain criteria and to be trained prior the evaluation task. Therefore, it is advisable to combine various metrics that evaluate different aspects in order to assure the reliability of the results.

In the light of this scenario, the aim of this paper is to describe the process carried out to develop a parallel corpus (Spanish – English) comprised of texts extracted from corporate websites of southern Spanish SMEs from the sanitary sector which will serve as the basis for future MT quality assessment tasks.

2 Methodology

Prior research set the basis for the corpus presented in this paper [14]. The latter study analyzed 1425 Andalusian SMEs, belonging to what is referred to as Group Q: Healthcare and social services activities according to the CNAE-2009 classification (Spanish National Classification of Economic Activities). This economic sector was chosen as the healthcare sector is the second biggest group as regards business creation in net terms according to official reports [15]. The sample was selected using information from the Sectoral Ranking of Companies by Turnover offered by the Spanish source elEconomista.es, a daily newspaper with special focus on economics, finance, and business. The data from this Company Ranking comes from the INFORMA D&B S.A.U. (S.M.E.) database − which boasts the Spanish Association for Standardization and Certification (AENOR) quality certificate − and is fed from several public and private sources. This study concluded that around a half of the analyzed SMEs had a website, but only 10% of them offered their content translated to one or more languages.

The final goal of the research project is to evaluate MT applied to corporate information available on SMEs websites, hence, reference translations were needed to build the parallel corpus. Therefore, those companies offering translated content to English and Spanish served to build the corpus described in this paper, which responds to a sequential sampling strategy meaning that first phase results [14] determined the methodology of the next phase [16, 17].

The stages for compiling the parallel corpora were: (i) selection of websites with content translated in English and Spanish, (ii) downloading of the HTML files of the selected websites, (iii) files filtering and pairing of English files with their Spanish equivalents, (iv) compilation of individual corpora (EN and ES) for each of the selected websites, (v) merging of the individual corpora into a two general corpus one in English and the other in Spanish, (vi) selection a representative sample of segments to be used as original (ES) and reference translations (EN) and, (vii) building of the parallel corpus intended for MT evaluation.

2.1 Selection of the sample

Previous research [14] showed that 64 companies offered their contents translated into English from Spanish. A technique of stratified random sampling [18, 19] was applied for selecting the websites which will comprise the corpus. Medical specialties were considered as the base for weighting adjustment. Therefore, the final sample was comprised of 45 websites (Table 1).

Table 1. Websites sample selection

Medical specialties	N	%	Sample (N)
Polyclinics and hospitals	7	10,94	5
Plastic Surgery	8	12,50	5
Radiology-Diagnostic	5	7,81	3

Obstetrics and Gynecology	9	14,06	6
Ophthalmology	1	1,56	1
Physical Medicine & Rehabilitation	4	6,25	3
Dentistry	19	29,69	12
Healthcare Transport	1	1,56	1
Oncology	1	1,56	1
Cardiology	1	1,56	1
Gastroenterology	1	1,56	1
Psychology	1	1,56	1
Otorhinolaryngology	1	1,56	1
Surgery	2	3,13	1
Neuroscience research	1	1,56	1
Neurology	1	1,56	1
Addiction treatment	1	1,56	1
Total	**64**	**100**	**45**

After selecting the 45 websites which will comprise the sample a professional scientific English native translator certified that all selected websites met the quality standards of professional translation

2.2 Downloading of websites

Once the websites were selected, they were downloaded with Cyotek Webcopy tool. A website is made of great volumes of files, that is why the download was limited to the first three depth levels. The reason behind this decision is that it is usually recommended placing the most relevant information in the first levels so that users do not have to click several times to access it and three levels are sufficient to meet this requirement [20, 21]. In total, 3.31 GB were downloaded, comprising 52,734 files and 15,741 folders.

2.3 Filtering of files and pairing

The downloaded files were filtered, and the HTML English files were paired to their equivalents in Spanish so that it was possible to obtain the reference translations, being Spanish the source text and English the target text. To this end, the files were named and stored to facilitate their identification. Two folders –English and Spanish−were created for each website and the files were named so that the Spanish version of the homepage of a given website was named as *Web1* and stored in the folder *Spanish* while its equivalent in English–named *Web1* as well–was stored in the folder *English*. Once all the files were paired those files not being useful for corpus compilation (HTML files without English equivalent, JavaScript files, etc.) were deleted.

2.4 Compilation of individual corpora for each website

Two corpora were compiled for each of the websites with Sketch Engine corpora analysis tool [22]. One of the corpora was built using the English files and the other using the Spanish files and once compiled, the resulting TXT files were downloaded. In total, 90 TXT files were obtained, half of them in English and the other half in Spanish.

2.5 Compilation of the general corpora

In order to know how many translation segments–sentences–were in total in the sample two general monolingual corpora, one in Spanish and the other in English, were built using the TXT files obtained in the previous stage. Table 2 show corpora description. The difference in the volume of tokens, words, sentences and paragraphs, besides the linguistic features of each language, is due to the fact that some of the Spanish files contained more information that their equivalents in English**.**

Table 2. General corpora description

	ES_Health SMEs websites	*EN_Health SMEs websites*
Tokens	726,093	613,524
Words	638,202	536,226
Sentences	43,450	38,053
Paragraphs	29,514	24,581

2.6 Sample selection

The sample selection process started once the two monolingual corpora were built. Given the corpus purpose–serve as the basis to perform MT evaluation−it was determined that sentences will be the reference unit to select translation segments. The number of sentences of the English corpus, 38,053, was established as a reference to estimate sample size. The sample was calculated for a finite population (N = 38,053) for a confidence level of 99% with margin of error of 5% [23]. Thus, the final sample was comprised of 654 segments.

Given the variability of the size and length of the companies websites, the segments extracted from each website ranged from 4,907−largest website–to 23–smallest website−. For his reason, a technique of stratifed random sampling [20, 21] was applied again for selecting the segments which will form the parallel corpus. In this case, the amount of segments of each website was considered as the base for weighting adjustment. It is important to note that two websites did not have sufficient percentual weight with regard to the total population, so they were not supposed to add any segment to the sample. However, in order not to leave two companies without representation a translation segment from each website was selected. As a result, the final sample was comprised of 656 segments (Table 3).

6

Table 3. Segments table selection

Company ID	Segments (N)	%	Sample (N)
8622_MAM10	4907	12,90	84
8690_MAM14	4321	11,36	74
8623_MAM08	2403	6,31	41
8690_MAP55	2134	5,61	37
8622_COM06	1843	4,84	32
8690_MAP42	1620	4,26	28
8622_MAP01	1613	4,24	28
8623_GRP10	1280	3,36	22
8690_CAP03	1250	3,28	21
8622_MAM06	1185	3,11	20
8610_CAP02	1123	2,95	19
8610_MAM08	1083	2,85	19
8690_MAM02	1067	2,80	18
8610_SEM01	1027	2,70	18
8622_SEM26	979	2,57	17
8690_MAP72	909	2,39	16
8621_MAP17	807	2,12	14
8610_MAM06	687	1,81	12
8690_CAM04	679	1,78	12
8610_MAP04	584	1,53	10
8622_MAP55	565	1,48	10
8623_MAP80	531	1,40	9
8622_MAP27	456	1,20	8
8621_MAP03	445	1,17	8
8622_MAP17	422	1,11	7
8622_ALP08	401	1,05	7
8623_SEP23	376	0,99	6
8621_ALP07	373	0,98	6
8623_ALM01	358	0,94	6
8623_MAP16	344	0,90	6
8622_MAP18	324	0,85	6
8622_CAM09	323	0,85	6
8623_MAP14	293	0,77	5
8623_MAP27	209	0,55	4
8690_COP10	194	0,51	3

8622_MAP22	189	0,50	3
8690_MAP47	172	0,45	3
8621_CAP16	134	0,35	2
8622_MAP47	112	0,29	2
8690_JAP09	101	0,27	2
8623_MAP88	75	0,20	1
8623_CAP11	61	0,16	1
8622_MAP28	49	0,13	1
8623_MAP52*	23	0,06	0 + 1
8690_CAP19*	22	0,06	0 + 1
TOTAL	**38,053**	**100**	**656**

3 Conclusion and future research

The aim of this paper was to describe step by step the methodological approach followed to build a parallel corpus which will be used for MT quality evaluation − including both human and automatic assessment – of corporate websites belonging to SMEs from the healthcare sector. To this end, to the final XLS file, containing the original Spanish segments and their equivalents in English, one more column was added containing MT output, therefore, the XLS file contains automatic translations (EN) together with their original translation (ES) and reference translations (EN), which are both essentials to MT quality assessment. However, more columns could be added in the future to include automatically generated translations from more MT systems to compare their performance. Further analysis concerning MT error identification, annotation and classification could also be carried out using the parallel corpus as a base, along with the evaluation of the post-editing process. In addition, the parallel corpora can be easily enlarged by adding segments from the monolingual corpora, which are already formatted and numbered in order to make the process as efficient as possible. The parallel corpus can also be uploaded to corpus analysis tools such as Sketch Engine for further linguistic analysis.

On another note, the monolingual corpora generated in English and Spanish can also serve to carry out linguistic research, including comparison between languages or monolingual analysis. These two corpora, given its considerable volume, could be used to train purpose-built MT systems and they can also serve to enlarge the knowledge concerning the features of corporate texts from the healthcare sector.

Acknowledgements

Study supported by the Spanish Ministry of Science, Innovation and Universities (MCIU), the State Research Agency (AEI) and the European Regional Development

Fund (ERDF) via the RTI2018.093348.B.I00 project and by the MCIU via the University Staff Training Program (FPU17/00667).

References

1. Alcaide, J.C., Bernués, S., Díaz-Aroca, E., Espinosa, R., Muñiz, R., Smith, C.: Marketing y Pymes. Las principales claves de marketing en la pequeña y mediana empresa (2013).
2. Muller, P., Julius, J., Herr, D., Koch, L., Peucheva, V., McKiernan, S.: Annual Report on European SMEs 2016/2017: Focus on self employment. Unión Europea, Bruselas (2017).
3. Eurostat: Internet advertising of businesses-statistics on usage of ads (2019).
4. Castilho, S., Doherty, S., Gaspari, F., Moorkens, J.: Approaches to Human and Machine Translation Quality Assessment. In: Moorkens, J., Castilho, S., Gaspari, F., and Doherty, S. (eds.) Translation Quality Assessment. pp. 9–38. Springer, Cham 2018.
5. Shaw, F., Gros, X.: Survey of Machine Translation Evaluation. , Saarbrücken (2007).
6. Han, L.: Machine Translation Evaluation Resources and Methods: A Survey. arXiv Comput. Lang (2016).
7. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions and reversals. Sov. Phys. Dokl. 10, 707–710 (1966).
8. Chatzikoumi, E.: How to evaluate machine translation: A review of automated and human metrics. Nat. Lang. Eng. 26, 137–161 (2020).
9. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Stroudsburg, PA, USA (2002).
10. Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In: HLT '02: Proceedings of the second international conference on Human Language Technology. pp. 138–144 (2002).
11. Lin, C.-Y., Hovy, E.: Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics. In: Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. pp. 150–157 (2003).
12. Turian, J.P., Shen, L., Melamed, I.D., Melamed, I.D.: Evaluation of Machine Translation and its Evaluation. In: Proceedings of MT Summit IX. , New Orleans (2003).
13. Banerjee, S., Lavie, A.: 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 Machine Translation and/or Summarization. pp. 65–72. , Ann Arbor (2005).
14. Rivera-Trigueros, I., Olvera-Lobo, M.D.: Internet Presence and Multilingual Dissemination in Corporate Websites: A Portrait of Spanish Healthcare SMEs. J. Glob. Inf. Manag. 29, 1–17 (2021).
15. Dirección General de Industria y de la Pequeña y Mediana Empresa: Retrato de la Pyme (2020).

16. Hernández-Sampieri, R., Fernández Collado, C., Baptista Lucio, P.: Metodología de la Investigación. McGrah-Hill (1991).
17. Baltar, F., Gorjup, M.T.: Muestreo mixto online: Una aplicación en poblaciones ocultas. Intang. Cap. 8, 123–149 (2012).
18. Babbie, E.: Fundamentos de la investigación social. Ediciones Paraninfo, Méjico (2000).
19. Clairin, R., Brion, P.: Manual de muestreo. La Muralla, Madrid (2001).
20. US Department of Health and Human Services: Research-Based Web Design & Usability Guidelines. US Government Printing Office, Washington, DC (2006).
21. Powell, T.A.: Web Design: The Complete Reference. , Berkley, CA (2000).
22. Kilgarriff, A., Rychlý, P., Smrz, P., Tugwell, D.: The Sketch Engine. In: Williams, G. and Vessier, S. (eds.) Proceedings of the 11th EURALEX International Congress. pp. 105–115. Lorient, Francia (2004).
23. Martínez Bencardino, C.: Estadística y muestreo. Ediciones ECOE, Bogotá (2019).