gApp: a text preprocessing system to improve the neural machine translation of discontinuous multiword expressions

Carlos Manuel Hidalgo-Ternero
Universidad de Málaga
Avda. Cervantes, 2. 29071
Malaga, Spain
cmhidalgo@uma.es

Xiaoqing Zhou-Lian
Universidad Rey Juan Carlos
Paseo de los Artilleros s/n. 28032
Madrid, Spain
xiaoqing.zhou@urjc.es

Abstract

In this paper we present research results with *gApp*, a text-preprocessing system designed for automatically detecting and converting discontinuous multiword expressions (MWEs) into their continuous forms so as to improve the performance of current neural machine translation systems (NMT) (see Hidalgo-Ternero, 2021 & 2022, Hidalgo-Ternero & Corpas Pastor, 2020, 2022a & 2022b, Hidalgo-Ternero, Lista, and Corpas Pastor, 2022, and Hidalgo-Ternero and Zhou-Lian, 2022a & 2022b). To test its effectiveness, eight experiments with several NMT systems such as DeepL, Google Translate, ModernMT and VIP have been carried out in different language directionalities (ES/FR/IT > ES/EN/DE/FR/IT/PT/ZH) for the translation of somatisms, i.e., MWEs containing lexemes referring to human or animal body parts (Mellado Blanco, 2004). More specifically, we have analysed both flexible verb-noun idiomatic constructions (VNICs) and flexible verb + prepositional phrase (VPP) constructions. In this regard, the promising results obtained for these typologies of MWEs throughout experiments 1-8 will shed some light on new avenues for enhancing MWE-aware NMT systems.

Introduction

The recent emergence of neural networks in natural language processing has represented a real breakthrough in the field of machine translation, bringing forth Neural Machine Translation (NMT), which has resulted in a considerable qualitative leap compared to previous ruled-based and statistical models (Bentivogli et al., 2016; Junczys-Dowmunt et al., 2016; Shterionov et al., 2018). Despite these advances, NMT systems still have an important weak spot: multiword expressions (MWEs). As well as their quintessential problematic features such as syntactic anomaly, non-compositionality, diasystematic variation and ambiguity, among others, a further challenge arises for NMT: MWEs do not always consist of adjacent tokens (e.g., *He took all their remarks into consideration.*), which seriously hinders their automatic detection and translation (Corpas Pastor, 2013; Foufi et al., 2019; Monti et al., 2018; Ramisch & Villavicencio, 2018; Rohanian et al., 2019). To overcome the challenges that discontinuous MWEs still pose for even the most robust NMT systems (cf. Colson, 2019; Zaninello & Birch, 2020), we have designed *gApp,*[1] a text-preprocessing system for the automatic identification and conversion of discontinuous MWEs into their continuous form in order to improve NMT performance. In this regard, 8 experiments, summarised in the *Results* section, have been carried out to prove *gApp*'s effectiveness.

Against this background, the remainder of the paper is structured as follows. Section 2 illustrates the research methodology. In Section 3, *gApp*'s precision and recall from experiments 1-8 is tested, in order to then assess to what extent this system can enhance NMT performance under the challenge of MWE discontinuity. Finally, Section 4 provides concluding remarks on how to further enhance MWE-aware NMT systems through *gApp*.

Methodology

This section presents the research methodology employed in order to assess to what extent *gApp* can optimise the performance of the NMT systems of DeepL, Google Translate, ModernMT and VIP in different language directionalities (ES/FR/IT > ES/EN/DE/FR/IT/PT/ZH). Analogously to Hidalgo-Ternero (2020), the concordances containing the discontinuous somatisms under study have been retrieved from two giga-token web-crawled corpus families (TenTen and Timestamped JSI web corpus) and the subcorpora available for the different languages under study (esTenTen18 and Timestamped JSI web corpus 2014-2021 Spanish, for Spanish; enTenTen20 and Timestamped JSI web corpus 2014-2021 English, for English, etc.). All these corpora are accessible through the corpus management system Sketch Engine (Kilgarriff et al., 2004).

The MWEs analysed through experiments 1-8 belong to the category of idiomatic expressions, since they have a non-compositional meaning (which is why they are also defined as semantically non-decomposable idioms or SNDIs [Bargman & Sailer, 2018]). Concerning their fixedness, following Parra et al.'s (2018) taxonomy for MWEs in Spanish, they can be classified as flexible, since other elements can appear embedded within the constituents of the MWEs. With regards to their morphosyntactic structure, they belong to two main categories: verb-noun idiomatic constructions (VNICs) and verb + prepositional phrase (VPP) constructions. Finally, considering the nature of their constituents, they are somatisms, i.e., idioms containing terms that refer to human or animal body parts (Mellado Blanco, 2004). In this regard, we have decided to analyse specifically idiomatic expressions because their non-compositional meaning makes them become potentially easier to detect and translate by NMT systems when all the constituents are contiguous, as we proved in experiments 1-8 (see Table 1 in the *Results* section).

Despite the challenges that user-generated content's (UGC) ubiquitous source-text error, noise and out-of-vocabulary tokens still pose to even the most robust NMT systems (Belinkov & Bisk, 2018; Lohar et al., 2019), a heterogeneous sample in terms of language varieties, text sources and types (including UGC) was selected for the

[1] *gApp* is accessible through the following link: http://lexytrad.es/gapp/app.php . This application is registered in Safecreative: https://www.safecreative.org/work/2011165898461-gapp.

analysis so as to alleviate sampling bias, which could otherwise originate from uniquely examining NMT canonical training data for the somatisms under study. In this way, a total of 3360 cases was analysed, comprising 1680 discontinuous and 1680 continuous forms (i.e., after the conversion) of somatisms, split by different unigrams, bigrams or trigrams. Besides these relevant results, for each somatism 50 irrelevant results (i.e., concordances containing analogous patterns to the MWEs but unrelated to the idiomatic sequences) were compiled, in order to calculate, at a first stage, both the precision and recall of this system, considering all the constituents of the MWE.

Once both parameters were quantified, at a second stage, the results concerning the NMT performance for the different concordances were classified within three main categories: before *gApp*, after the automatic conversion with *gApp*, and after the manual conversion, which hence constituted our gold standard. The same study was conducted for all the language directionalities. The NMT outputs for these different scenarios were then manually assessed following an instance-based MT evaluation (Zaninello & Birch, 2020) with several possible target-text candidates for each of the somatisms in both their continuous and discontinuous forms. To this end, morphological, syntactic, and/or orthotypographic divergences or source-text/translation imprecisions affecting other elements in the sentences were not considered *per se* as errors if they were unrelated to the phenomenon of MWE discontinuity for the somatisms under study.

Results

Eight different experiments, summarised in Table 1, have been carried out to prove *gApp*'s effectiveness.

	Type of MWE	Language directionalities	NMTs analysed	NMTs' accuracy before *gApp*	NMTs' accuracy after *gApp*	Improvement after *gApp*	Manual conversion
1.	VNC	ES>EN	DeepL	80.7%	90.7%	10%	+3.2%
			Google Translate	60.7%	75.4%	14.6%	+2.1%
2.	VNC	ES>EN	DeepL	49%	62.5%	13.5%	+0.5%
		ES>DE		43.5%	52.5%	9%	+0.5%
3.	VNC	FR>EN	DeepL	40%	58%	18%	=
		FR>ES		41.5%	58%	16.5%	=
4.	VPP	ES>EN	Modern MT	50%	60%	10%	=
		ES>DE		23.3%	33.3%	10%	=
		ES>FR		49.3%	60%	10.7%	=
		ES>IT		56.7%	60.7%	4%	=
		ES>PT		56%	58.7%	2.7%	=
		ES>EN	DeepL	70.7%	81.3%	10.7%	+0.7%
		ES>DE		59.3%	66.7%	7.3%	+0.7%
		ES>FR		69.3%	74%	4.7%	+0.7%
		ES>IT		76%	80%	4%	+0.7%

		ES>PT		68%	74%	6%	+0.7%
		ES>EN	Google Translate	66%	75.3%	9.3%	=
		ES>DE		35.3%	43.3%	8%	=
		ES>FR		65.3%	73.3%	8%	=
		ES>IT		78.7%	79.3%	0.7%	=
		ES>PT		72.7%	79.3%	6.7%	=
5.	VPP	ES>EN	VIP	45.5%	67%	21.5%	-0.5%
		ES>EN	DeepL	77%	85.5%	8.5%	=
		ES>EN	Google Translate	64%	77.5%	13.5%	-1%
6.	VPP	IT>EN	Modern MT	25.5%	42%	16.5%	+1%
		IT>DE		28%	37%	9%	=
		IT>EN	Google Translate	64.5%	82.5%	18%	+0.5%
		IT>DE		50.5%	61%	10.5%	-1%
		IT>EN	DeepL	75%	75%	0%	-0.5%
		IT>DE		62%	67.5%	5.5%	=
7.	VNC	ES>EN	Google Translate	21.5%	25%	3.5%	=
			DeepL	57%	54.5%	-2.5%	=
		ES>ZH	Google Translate	11%	14%	3%	=
			DeepL	42.5%	39.5%	-3%	=
8.	VPP	ES>EN	Google Translate	13.6%	66.0%	52.4%	+0.8%
			DeepL	66%	90%	24%	=
		ES>ZH	Google Translate	13.6%	64.8%	51.2%	+0.8%
			DeepL	54.8%	73.6%	18.8%	=
Total (experiments 1-8)				**52.1%**	**66.8%**	**14.6%**	**+0.5%**

Table 1. *gApp* results through experiments 1-8

In Table 1, it is possible to observe a considerable improvement in NMT performance from 52.1% before *gApp* up to 66.8% after *gApp* (i.e., a final enhancement by 14.6%). Global results have also shown how *gApp*'s automatic conversion managed to achieve an analogous performance to the manual conversion (with only a 0.5% difference between the two types of conversion). This is chiefly due to *gApp*'s refined detection system both in terms of final average precision (95.9%) and recall (97.3%), which means that only 4.1% of the irrelevant results could enter the system and exclusively 2.7% of the relevant results were not successfully detected. A summary of *gApp*'s precision and recall through experiments 1-8 can be observed in Table 2.

	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6	Exp. 7	Exp. 8	Global
Precision	94.8%	95.1%	96.1%	94.9%	95.2%	98.3%	95.3%	97.3%	95.9%
Recall	96.8%	97.5%	98.5%	99.3%	96%	92%	99.5%	98.8%	97.3%
F_1	95.8%	96.3%	97.3%	97.1%	95.6%	95.2%	97.4%	98.1%	96.6%

Table 2. *gApp* precision and recall through experiments 1-8

Other interesting findings can be observed in target-text errors in different language directionalities due to NMT pivoting through English. In this regard, let us contrast some instances of DeepL's performance for the Spanish somatisms *bajar los brazos* and *arrimar el hombro* into English and German.

	KWIC extracts
ST [ES]	Las dificultades del primero para iniciar el juego colaboraron en alguno de los goles rivales; el segundo trato de dar coherencia al juego de un equipo horroroso en la transición defensiva, hasta que bajó los brazos definitivamente.
	DeepL's outcomes
TT [EN]	The difficulties of the first one to start the game collaborated in some of the rival goals; the second one tried to give coherence to the game of a horrendous team in the defensive transition, until it gave up the arms definitively.
TT [DE]	Die Schwierigkeiten des ersten, das Spiel zu beginnen, wirkten bei einigen der rivalisierenden Tore mit; der zweite versuchte, dem Spiel einer horrenden Mannschaft in der defensiven Übergangsphase Kohärenz zu verleihen, bis er die Waffen endgültig abgab.

Table 3. Instances of DeepL mistranslations in English and German for the somatism *bajar los brazos*

	KWIC extracts
ST [ES]	Por esta razón sólo cabía la posibilidad de arrimar el hombro un poco y realizar las aportaciones y modificaciones económicas necesarias, para conseguir una plaza de toros más viable.
	DeepL's outcomes
TT [EN]	For this reason, there was only the possibility of putting the shoulder to the wheel a little and making the necessary contributions and economic modifications, in order to achieve a more viable bullring.
TT [DE]	Aus diesem Grund gab es nur die Möglichkeit, ein wenig die Schulter ans Steuer zu legen und die notwendigen Beiträge und wirtschaftlichen Änderungen vorzunehmen, um eine lebensfähigere Stierkampfarena zu erreichen.

Table 4. Instance of DeepL mistranslation in German for the somatism *arrimar el hombro*

In Table 3 it is possible to observe that, in the ES>DE directionality, *bajar los brazos* has been translated as *die Waffen angeben* ('to give up the weapons'). The only way to understand what yielded this unpredictable outcome is to analyse the English target text for the source-text (ST) somatism in the ES>EN directionality: *to give up the arms*. Therefore, in the ES>DE scenario, the German version was mostly determined by the training data with English with a misinterpretation of *the arms* as *die Waffe* ('the weapons') instead of *die Arme* ('the arms' as body parts). Analogous mistranslations can be observed when examining DeepL's outcomes in other language directionalities for this Spanish ST idiom: *abandoner les armes* ('to abandon the weapons') in French, and *cedere le armi* ('to give in the weapons') in Italian. In Table 4, a similar problem can be detected. In this case, while in ES>EN an appropriate equivalent for the ST somatism *arrimar el hombro* has been offered (*to put the shoulder to the wheel*), in the ES>DE scenario it is possible to detect the sequence *die Schulter ans Steuer legen* ('to put the shoulders on the [steering] wheel'), with no idiomatic meaning. Once again, similar mistranslations with no idiomatic readings are to be found in other language directionalities for this Spanish ST somatism: *mettre l'épaule à la roue* in French, *mettere la spalla alla ruota* in Italian, or *colocar o ombro na roda* in Portuguese. These mistranslations hence emphasise the necessity for more training data in language combinations different from English, in order to avoid English-centred NMT outcomes.

Conclusion

The findings of our study confirm our hypothesis: the system *gApp* can, on average, improve the quality of the neural machine translation of discontinuous MWEs by converting them into their continuous form. More specifically, *gApp* has proved to enhance NMT for the analysed MWEs with a final average amelioration by 14.6%, which is only a 0.5% lower than the gold standard (15.1%).

These promising results with VNC and VPP somatisms in different language directionalities invite to further increase *gApp*'s detection lexicon and conversion mechanism so as to evaluate to what extent it can also result in NMT enhancement for other discontinuous MWE categories. In addition, the present study can also constitute the basis for further research to assess the scalation of this model to other language-dependent text preprocessing systems for the automatic conversion of discontinuous MWEs in syntactically flexible languages, with the purpose of enhancing MWE-aware NMT systems.

Acknowledgements

This research has been carried out within the framework of several research projects (ref. PID2020-112818GB-I00, UMA18-FEDERJA-067, P20-00109, E3/04/21, UMA-CEIATECH-04 and 03/2021-Embassy of France in Spain) at Universidad de Málaga (Spain) and at Research Institute of Multilingual Language Technologies (IUITLM). It has also been funded by a post-doc grant entitled *Ayuda para la recualificación del Sistema Universitario Español 2021-2023 (Modalidad «Margarita Salas»)* at Universidad de Málaga and at Université catholique de Louvain (Belgium).

References

Bargmann, Sascha and Manfred Sailer. 2018. The syntactic flexibility of semantically non-decomposable idioms. In Manfred Sailer and Sascha Markantonatou (Eds.), *Multiword expressions: Insights from a multi-lingual perspective*, pages 1–29. Language Science Press.

Belinkov, Yonatan, and Yonatan Bisk. 2018. Synthetic and natural noise both break neural machine translation. *ArXiv*. https://arxiv.org/abs/1711.02173

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo, and Marcello Federico. 2016. Neural versus phrase-based machine translation quality: a case study. *arXiv*. preprint arXiv:1608.04631.

Colson, Jean-Pierre. 2019. Multi-word Units in Machine Translation: why the Tip of the Iceberg Remains Problematic – and a Tentative Corpus-driven Solution. *MUMTT2019*.

Corpas Pastor, Gloria 2013. Detección, descripción y contraste de las unidades fraseológicas mediante tecnologías lingüísticas. In Inés Olza and Elvira Manero (Eds.), *Fraseopragmática* (pages 335-373). Frank & Timme.

Foufi, Vasiliki, Luca Nerima and Eric Wehrli. 2019. Multilingual parsing and MWE detection. In Yannick Parmentier and Jakub Waszczuk (Eds.), *Representation and parsing of multiword expressions: Current trends* (pages 217–237). Language Science Press.

Hidalgo-Ternero, Carlos Manuel. 2020. Google Translate vs. DeepL: analysing neural machine translation performance under the challenge of phraseological variation. *MonTI. Monografías de Traducción e Interpretación, Special Issue 6*, 154-177. https://doi.org/10.6035/MonTI.2020.ne6.5

Hidalgo-Ternero, Carlos Manuel. 2021. El algoritmo ReGap para la mejora de la traducción automática neuronal de expresiones pluriverbales discontinuas (FR>EN/ES). In Gloria Corpas Pastor, María Rosario Bautista Zambrana and Carlos Manuel Hidalgo-Ternero (Eds.), *Sistemas fraseológicos en contraste: enfoques computacionales y de corpus*. Comares (pages 253-270)

Hidalgo-Ternero, Carlos Manuel. 2022/forthcoming. A la cabeza de la traducción automática neuronal asistida por gApp: somatismos en VIP, DeepL y Google Translate. In Gloria Corpas Pastor and Míriam Seghiri (Eds.). Comares.

Hidalgo-Ternero, Carlos Manuel, and Gloria Corpas Pastor. 2020. Bridging the 'gApp': improving neural machine translation systems for multiword expression detection. *Yearbook of Phraseology, 11*(1), 61–80. https://doi.org/10.1515/phras-2020-0005

Hidalgo-Ternero, Carlos Manuel, and Gloria Corpas Pastor. 2022a/forthcoming. ReGap: a text preprocessing algorithm to enhance MWE-aware neural machine translation systems. In Johanna Monti, Gloria Corpas Pastor and Ruslan Mitkov (Eds.), *Recent Advances in MWU in Machine Translation and Translation technology*. John Benjamins Publishing Company.

Hidalgo-Ternero, Carlos Manuel and Gloria Corpas Pastor. 2022b/forthcoming. Qué se traerá gApp entre manos… O cómo mejorar la traducción automática neuronal de variantes somáticas (ES>EN/DE/FR/IT/PT). In Míriam Seghiri and Míriam Pérez Carrasco (Eds.). *Aproximación a la traducción especializada*. Peter Lang.

Hidalgo-Ternero, Carlos Manuel, Francesco Lista, and Gloria Corpas Pastor. 2022/under review. gApp-assisted NMT: how to improve the neural machine translation of discontinuous multiword expressions (IT>EN/DE). *Language Resources and Evaluation*.

Hidalgo-Ternero, Carlos Manuel, and Xiaoqing Zhou-Lian. 2022a. Reassessing gApp: Does MWE Discontinuity Always Pose a Challenge to Neural Machine Translation?. In Gloria Corpas Pastor and Ruslan Mitkov (eds) *Computational and Corpus-Based Phraseology. EUROPHRAS 2022. Lecture Notes in Computer Science*, vol 13528. Springer, Cham. https://doi.org/10.1007/978-3-031-15925-1_9

Hidalgo-Ternero, Carlos Manuel, and Xiaoqing Zhou-Lian. 2022b/under review. Minding the gApp in the ES>EN/ZH neural machine translation of discontinuous multiword expressions. *Natural Language Engineering*

Junczys-Dowmunt, Marcin, Tomasz Dwojak, and Hieu Hoang. 2016. Is neural machine translation ready for deployment? A case study on 30 translation directions. *Arxiv*. https://arxiv.org/pdf/1610.01108.pdf

Kilgarriff, Adam, Pavel Rychly, Pavel Smrz, and David Tugwell. 2004. The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*, pages 105-116.

Lohar, Pintu, Maja Popović, Haithem Alfi, and Andy Way. 2019. A systematic comparison between SMT and NMT on translating user-generated content. *20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)*.

Mellado Blanco, Carmen. 2004. *Fraseologismos somáticos del alemán*. Peter Lang, Frankfurt.

Monti, Johanna, Violeta Seretan, Gloria Corpas Pastor and Ruslan Mitkov. 2018. Multiword units in machine translation and technology. In R. Mitkov, J. Monti, G. Corpas Pastor & V. Seretan (Eds.), *Multiword Units in Translation and Translation Technology*, pages 1-37. John Benjamins.

Parra Escartín, Carla, Almudena Nevado Llopis, and Eoghan Sánchez Martínez. 2018. Spanish multiword expressions: Looking for a taxonomy. In Manfred Sailer and Stella Markantonatou (eds.), *Multiword expressions: Insights from a multi-lingual perspective*, 271–323. Language Science Press.

Ramisch, Carlos and Aline Villavicencio. 2018. Computational treatment of multiword expressions. In Ruslan Mitkov (Ed.), Oxford Handbook on Computational Linguistics (2ª ed). https://doi.org/10.1093/oxfordhb/9780199573691.013.56

Rohanian, Omid, Shiva Taslimipoor, Samaneh Kouchaki, Le An Ha, and Ruslan Mitkov. 2019. Bridging the Gap: Attending to Discontinuity in Identification of Multiword Expressions. In Jill Burstein, Christy Doran, and Thamar Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies 1*. (pages 2692–2698). Association for Computational Linguistics.

Shterionov, Dimitar, Riccardo Superbo, Pat Nagle, Laura Casanellas, Tony O'Dowd, and Andy Way. 2018. Human versus automatic quality evaluation of NMT and PBSMT. *Machine Translation*, 32, 217–235. https://doi.org/10.1007/s10590-018-9220-z

Zaninello, Andrea and Alexandra Birch. 2020. Multiword expression aware neural machine translation. *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 3816–3825.