Universidade do Minho

Universidade de Vigo

linguaMATICA

Volume 16, Número 1 (2024)

ISSN: 1647-0818

Volume 16, Número 1 – 2024

LinguaMÁTICA

ISSN: 1647–0818

Conteúdo

Revisão

A comissão científica da **Linguamática** pode ser consultada na página Web da revista, em https://linguamatica.com/index.php/linguamatica/about/editorialTeam.

Para esta edição, colaboraram os seguintes investigadores:

- **Álvaro Iriarte Sanroman**, Universidade do Minho, Portugal
- **Eugénio Ribeiro**, Instituto Superior Técnico, Portugal
- **Gerardo Sierra**, Universidad Nacional Autónoma de México, México
- **Jorge Baptista**, Universidade do Algarve, Portugal
- **Lluís Padró**, Universitat Politècnica de Catalunya, Espanha
- **Miquel Esplà Gomis**, Universitat d'Alacant, Espanha
- **Rogelio Nazar**, Pontificia Universidad Católica de Valparaíso, Chile

Artigos de Investigação

Proposta rebuda el de desembre 2023 i acceptada per a publicació el de juny 2024.

Resolució anafòrica en traducció automàtica: el cas de l'espanyol i el català

Anaphoric Resolution in Machine Translation: the Case of Spanish and Catalan

Sergi Alvarez-Vidal
Universitat Pompeu Fabra

Resum

En l'última dècada, la traducció automàtica (TA) ha augmentat la seva presència no només en el sector de la traducció sinó també en el conjunt de la societat, en part pels bons resultats de qualitat obtinguts per la traducció automàtica neuronal (TAN). Actualment, els models massius de llenguatge (MML) com ara GPT (Generic Pre-trained Transformer) poden generar text sobre una infinitat de temes diferents i també traduir documents tenint en compte un context més ampli. Tot i així, per a idiomes estretament relacionats, com ara l'espanyol i el català, la traducció automàtica basada en regles (TABR) s'utilitza diàriament per traduir milers de paraules.

Aquest article estudia la TAN, TABR i GPT del castellà al català, dues llengües romàniques amb una estructura molt semblant en les quals els sistemes de TABR han demostrat un bon rendiment. Utilitzem un *challenge test set* centrat en la resolució d'anàfores, específicament els pronoms febles, un grup de pronoms que no tenen una correlació directa entre les dues llengües. Com que els models de TABR només tenen en compte la informació a nivell de frase, només estudiem les aparicions intraoracionals. L'objectiu és avaluar un fenomen sintàctic complex que ens pot ajudar a entendre quin dels tres sistemes tradueix més bé els elements contextuals.

Els resultats mostren que els dos models GPT provats són els que produeixen el nombre més baix d'errors, seguit dels sistemes de TAN. Tot i així, el nombre de traduccions errònies en el millor sistema és del 47%, cosa que contrasta amb els bons resultats d'avaluació generals que s'obtenen per a aquest parell de llengües.

Paraules clau

traducció automàtica, TA, MML, GPT, traducció automàtica basada en regles, traducció automàtica neuronal, anàfora, pronoms febles

Abstract

In the last decade, machine translation (MT) has increased its presence not only in the translation industry but also in society as a whole, in part due to the good results in quality produced by neural machine translation (NMT). Currently, large language models (LLMs) such as GPT (Generic Pre-trained Transformer) can generate text on endless topics, and also translate documents taking into account a larger context. Even so, for closely-related languages such as Spanish and Catalan rule-based machine translation (RBMT) is used daily to translate thousands of words.

This article studies how RBMT, NMT and GPT perform translating from Spanish into Catalan, two Romance languages with very similar structure in which RBMT systems have shown to perform well. We use a challenge test set focusing on anaphora resolution, specifically weak pronouns, a group of pronouns which do not have a direct correlation between the two languages. As RBMT models only take into account sentence level information, we only study intra-sentential appearances. The goal is to assess a complex syntactic phenomenon which can help understand which system translates better contextual information.

Results show the two GPT models tested are the ones with the less number of errors, followed by the NMT models. Even so, the number of errors in the model with the best results is 47%, which does not correspond to general assessment results usually obtained for this language combination.

Keywords

machine translation, MT, LLM, GPT, rule-based machine translation, neural machine translation, anaphora, weak pronouns

1. Introducció

Fa dècades que es fa servir la traducció automàtica com a ajuda a la traducció. En un principi, es van començar a fer servir sistemes de traducció automàtica basats en regles (TABR). Aquests sistemes requereixen un diccionari i una sèrie de regles que permeten passar una frase d'una llengua A a una llengua B. Amb el pas dels anys, aquests sistemes van incorporar una anàlisi gramatical inicial per fer la traducció d'estructures en lloc de paraules simples.

DOI: 10.21814/lm.16.1.424

Més endavant, molts d'aquests sistemes es van anar substituint per a molts parells de llengües per sistemes basats en corpus, com la traducció automàtica estadística (TAE). Aquests sistemes exigeixen una capacitat de computació molt més gran i trien la traducció per a cada frase en funció de les probabilitats extretes a partir d'un corpus d'entrenament. En els últims anys, i gràcies a l'augment exponencial de les capacitats computacionals, ha aparegut la traducció automàtica neuronal (TAN), que està basada en xarxes neuronals artificials.

Aquest nou model ha obtingut molt bons resultats de qualitat (Vaswani et al., 2017; Bahdanau et al., 2014) i això ha fet que la presència de la traducció automàtica s'hagi multiplicat en tots els àmbits, tant en la indústria de la traducció com en moltes altres situacions de comunicació. En el cas del castellà i el català, que són dues llengües romàniques amb moltes similituds morfològiques, semàntiques i sintàctiques, fa anys que es fan servir sistemes de traducció automàtica per regles per traduir milers de paraules diàries en diferents àmbits (Fité Labaila, 2007).

Amb l'actual popularitat dels sistemes de TAN, moltes empreses i institucions estan substituint progressivament els sistemes que fan servir per nous sistemes de TAN. Tot i que la recerca mostra un increment de la qualitat (especialment de la fluïdesa del text d'arribada) per als sistemes neuronals (Castilho et al., 2017), cal que l'avaluació de la qualitat tingui en compte el parell concret de llengües de treball. En les llengües properes amb estructures sintàctiques similars, s'ha mostrat una lleugera millora en l'avaluació automàtica i manual per als sistemes TAN en el cas del castellà i el català (Alvarez et al., 2019), però no en el cas del castellà i el gallec (Do Campo Bayón & Sánchez-Gijón, 2019).

A banda d'això, acaben d'aparèixer Generative Pre-trained Transformers (GPT) com ara ChatGPT, que són un tipus de model massiu de llenguatge (MML). En principi són models generatius multilingües dissenyats per contestar preguntes i parlar sobre gairebé qualsevol tema, tot i que també permeten traduir tenint en compte un context més ampli (Castilho et al., 2023). Això fa que es plantegi la pregunta de quin d'aquests sistemes té realment un millor rendiment i qualitat per al cas de la traducció del castellà al català.

Ja des dels primers models de traducció automàtica, la traducció dels pronoms va resultar un repte important (Hobbs, 1978). Això es deu principalment a l'ambigüitat d'interpretació que poden comportar determinats pronoms i a la necessitat de tenir informació semàntica i contextual addicional per poder resoldre la referència anafòrica. Actualment encara és un problema recurrent per a tots els sistemes de TA. Alguns investigadors argumenten que aquestes dificultats es deuen al fet que la TA parteix de la frase com a paradigma bàsic (Wicks & Post, 2022). Això ha fet que l'avaluació de la TA sigui cada cop més conscient que cal avaluar tot el document (Barrault et al., 2020).

En aquest article analitzem la resolució d'anàfores en la traducció del castellà al català. Concretament, ens centrem en els pronoms febles. Aquests pronoms presenten grans diferències d'ús entre les dues llengües i això ens permet estudiar com es resolen. Hem decidit limitar l'ús dels pronoms febles a contextos interoracionals perquè els tres models que estudiem (TABR, TAN i MML) treballin en igualtat d'oportunitats. Per fer-ho, hem creat un *challenge test set* o *test suite*, un conjunt de frases creades *ad hoc* amb combinacions complexes de pronoms febles que permetin posar a prova la capacitat de traducció dels diferents sistemes.

2. TA neuronal, per regles i GPT

La traducció automàtica basada en regles (TABR) és un sistema que modifica el text original a partir de diferents regles gramaticals i lèxiques per traduir el text a la llengua d'arribada. Això suposa un procés llarg de creació de regles per a la transferència però també permet controlar, modificar i afinar totes les regles que s'apliquen al llarg de la fase d'anàlisi, transferència i generació (España-Bonet et al., 2011). El principal problema rau en l'alt cost humà que implica, ja que cal codificar a mà totes les regles, que s'han d'anar afinant i modificant a mesura que apareixen errors o s'incorporen nous elements lèxics. Tanmateix, encara es fa servir per a llengües que tenen disponibles menys dades per a l'entrenament (Islam et al., 2022; Sghaier & Zrigui, 2020; Bayatli et al., 2018) o per a llengües properes amb semblances sintàctiques, com en el cas del castellà i el català. En un estudi recent en el qual es comparaven diferents sistemes de TA, el sistema de TABR va ser triat en el 31-43% dels casos (Aranberri et al., 2017). Un dels productes que actualment utilitza aquesta tecnologia és Apertium (Forcada et al., 2011).

Com que el català i el castellà són totes dues llengües oficials a Catalunya, hi ha una clara necessitat de generar traduccions perquè molts documents estiguin en les dues llengües, com en el cas del Diari Oficial de la Generalitat. Una

de les fites que va demostrar l'abast dels sistemes de traducció automàtica va ser l'aparició, el 28 d'octubre de 1997, de la primera la traducció diària que es va fer d'*El Periódico de Catalunya* al català. Aquest diari, publicat en castellà, va decidir publicar una versió diària en català amb l'ajuda d'un sistema de traducció automàtica i un grup d'uns 40 lingüistes. Al principi era un sistema simple de transferència per regles que incorporava un diccionari de paraules i seqüències i algunes regles morfològiques bàsiques que produïuen una traducció literal (Fité Labaila, 2001). Tanmateix, amb el pas del temps i l'ajuda dels lingüístes que feien esmenes i correccions diàries als resultats produïts per la TA, el sistema va anar millorant fins a implementar una nova versió el 2004 (Fité Labaila, 2007). Aquesta nova versió presentava regles morfològiques i sintàctiques molt més complexes i la TA produïa uns resultats força precisos.

La traducció automàtica neuronal (TAN) (Forcada, 2017). Necessita una gran quantitat de dades i fa servir xarxes neuronals. Com que ha millorat la qualitat dels resultats per a una gran varietat de llengües tant amb les evaluacions humanes com automàtiques (Castilho et al., 2017; Bentivogli et al., 2018), la majoria d'empreses han passat a utilitzar aquesta tecnologia per produir les seves traduccions.

La TAN s'ha comparat àmpliament amb els altres sistemes de traducció automàtica disponibles, especialment els sistemes de traducció automàtica estadística (TAE), que són els que es feien servir habitualment just abans de l'aparició dels sistemes neuronals, i que encara es fa servir en determinats àmbits. La majoria de comparacions destaquen la millora que es produeix en la qualitat, centrada sobretot en la fluïdesa de les frases que produeixen els sistemes neuronals (Castilho et al., 2017). Part de la recerca també s'ha centrat a analitzar les diferències entre els sistemes neuronals i els sistemes de TABR. Buysschaert et al. (2018) van fer servir un *challenge test set* (vegeu la secció 4) per estudiar quins eren els errors més freqüents que produïen els diferents sistemes de TA. Van veure que, tot i que la TAN obté millors resultats en la composició, les dependències de llarga distància, les expressions formades per diverses paraules i la subordinació, generaven més variació. En canvi, els sistemes de TABR resolien més bé els casos d'ambigüitat, i els sistemes de TAE obtenien millors resultats en la terminologia i els noms propis.

Brussel et al. (2018) van fer una anàlisi detallada dels errors produïts per diferents sistemes de traducció automàtica basats en regles, estadístics i neuronals en traduccions de l'anglès al neerlandès. Els autors van detectar que en termes generals la TAN produïa millors resultats, especialment si es tenia en compte la fluïdesa del text de destinació. Tanmateix, les millores en la precisió de les traduccions no eren tan clares. La TABR contenia el nombre més baix d'omissions i obtenia millors resultats en les oracions de més de 40 caràcters. La TAN només tenia uns resultats inferiors als altres dos sistemes per als errors relacionats amb tries lèxiques.

Koponen et al. (2019) van comparar els canvis introduïts en la postedició de documents traduïts de l'anglès al finès amb motors de TAN, TABR i TAE, i van dur a terme una anàlisi tant del producte final com del procés. Van arribar a la conclusió que la modificació més freqüent que s'introduïa en posteditar traduccions automàtiques basades en regles eren les eliminacions, però que gestionava millor la traducció de formes verbals i ambigüitats que els altres dos sistemes de traducció automàtica. La TAN mostrava un lleuger empitjorament en la mitjana de la longitud de pauses.

Actualment es fan servir els sistemes de TABR per traduir milers de paraules diàries entre el castellà i el català. Per exemple, el sistema de TABR desenvolupat per traduir *El Periódico de Catalunya* (Fité Labaila, 2007) es fa servir encara per traduir la versió catalana d'aquest diari i també *La vanguardia*. Malgrat això, la TAN ha mostrat una millora de la qualitat per a aquestes dues llengües (Alvarez et al., 2019; Costa-jussà, 2017) tot i que per a altres llengües d'estructura sintàctica propera no ha produït millores significatives (Do Campo Bayón & Sánchez-Gijón, 2019).

Els avenços recents en el processament del llenguatge natural (PLN) han impulsat el desenvolupament de models massius de llenguatge (MML), que han suposat millores notables en moltes tasques de processament del llenguatge. Tot i que aquests models multilingües s'han dissenyat per parlar amb els usuaris, han obtingut molt bons resultats quan s'han aplicat a tasques de traducció automàtica (Hendy et al., 2023). Entre aquests models, destaquen els models de Generative Pre-trained Transformer (GPT) (Brown et al., 2020), que han rebut molta atenció per la seva capacitat de generar textos coherents que tenen en compte el context i són capaços d'interactuar i modificar les respostes en funció de les preguntes o demandes concretes que se'ls plantegen.

Aquests models GPT i els sistemes de traducció automàtica neuronal (TAN) estan tots dos basats en l'arquitectura de transformer (Vaswani et al., 2017) però presenten certes diferències. Els models GPT són només de descodificador, i fan servir els mateixos paràmetres per processar el context i el text d'origen com una sola entrada per generar la propera sortida. A banda d'això, els models TAN normalment tenen una arquitectura de codificador-descodificador que codifica la frase d'origen en la xarxa del codificador i descodifica la frase de destinació tenint en compte les sortides anteriors. Els models GPT acostumen a estar entrenats només amb dades monolingües i, a més, necessiten una quantitat molt més gran de dades.

L'objectiu d'aquest estudi és comparar dos sistemes de TABR, dos de TAN i dos models GPT per a la traducció del castellà al català. Tanmateix, no s'avalua la qualitat general, sinó la traducció dels pronoms febles, un fenomen complex de referència anafòrica que sovint presenta dificultats de traducció i necessita tenir en compte l'antecedent per poder resoldre's.

3. Anàfora i pronoms febles

Una anàfora és un element lingüístic que fa referència a un altre element lingüístic que s'esmenta en el text (Tognini-Bonelli, 2001). L'element al qual fa referència és l'antecedent i, per definició, depèn d'aquest element per a la seva interpretació (van Deemter & Kibble, 2000). L'anàfora, doncs, ens permet recuperar elements que ja han estat presentats sense haver de repetir els mateixos conceptes i és un recurs habitual en la majoria de textos.

Des que es van començar a desenvolupar sistemes de TA, es va detectar la dificultat que suposa la resolució de les anàfores. Hobbs (1978) ho il·lustrava amb un exemple molt clar:

- There is a pile of inflammable trash next to your car. You'll have to get rid of it.

De fet, l'humor que genera aquesta frase d'un episodi d'una coneguda sèrie de televisió nordamericana es deu precisament a la possible doble interpretació del pronom a la segona frase. Tanmateix, nosaltres entenem ràpidament quin és el sentit primer de la frase perquè tenim en compte el context.

A banda de la dificultat pròpia de l'anàfora, la traducció hi afegeix una complexitat addicional. Després que el receptor identifiqui i descodifiqui l'antecedent de l'anàfora consignada per l'emissor, l'ha de tornar a codificar en una altra llengua.

La TA sempre ha considerat la resolució d'anàfores com un repte, però en funció dels diferents models s'han adoptat diferents estratègies per resoldre-la. Mitkov (1999) lamentava la poca atenció que rebia aquest problema i la seva recerca afegeix característiques addicionals per a la resolució d'anàfores a la representació intermèdia d'un model de transferència (Mitkov et al., 1995). Lappin & Leass (1994) estableixen una sèrie de regles ordenades que cal aplicar a un sistema de TA per resoldre l'ús dels pronoms.

Una gran part dels primers models de resolució d'anàfores estaven basats en l'anàlisi sintàctica. Hobbs (1978) planteja un estudi de l'anàlisi sintàctica de les frases que permet establir quin és l'antecedent que es prioritza. Lappin & Leass (1994) també apliquen l'anàlisi sintàctica per indicar els possibles antecedents i, després, apliquen un sistema de pesos per decidir quin és l'antecedent més probable.

En aquest camp també s'ha fet servir l'aprenentatge automàtic. Ge et al. (1998) presenten un algoritme basat en l'estudi de dades estadístiques. Kehler et al. (2004), en canvi, extreuen probabilitats de referència a partir d'un corpus anotat.

La irrupció dels models neuronals i, més recentment, dels models massius de llenguatge, ha fet que ens puguem plantejar superar el paradigma tradicional de les frases com a element de treball (Wicks & Post, 2022) i hi incloguen el context. Això, en principi, hauria de ser beneficiós per a la qualitat de la traducció, especialment per als fenòmens de coherència textual i ambiguïtat, entre els quals s'inclou l'anàfora. De fet, la recerca en aquesta sentit demostra que la inclusió del context millora la resolució anafòrica (Voita et al., 2018; Castilho et al., 2023).

L'anàfora, doncs, es pot produir en dos contextos diferents: pot fer referència a un antecedent dintre de la mateixa frase (intraoracional) o a un antecedent en una frase anterior (interoracional). Per al nostre *challenge test set*, com hem comentat abans, només farem servir oracions intraoracionals, ja que els sistemes de TABR no són capaços d'ampliar el context que tenen en compte.

Hi ha diferents tipus d'anàfora que poden implicar, entre altres, pronoms, demostratius, elements nominals i el·lipsis. Un dels elements anafòrics més habituals són els pronoms. En català, tenim pronoms forts i febles. Els pronoms forts són "pronoms personals amb accent de mot que poden ocupar qualsevol de les posicions sintàctiques d'un sintagma nominal" (Institut d'Estudis Catalans, 2016). Els pronoms febles

són "pronoms sense accent de mot que s'anteposen o es posposen al verb i formen amb aquest una unitat accentual"(Institut d'Estudis Catalans, 2016).

Aquest article se centra en l'ús dels pronoms febles en les traduccions del castellà al català, perquè en molts casos no hi ha una correlació directa en les traduccions entre aquestes dues llengües. Per fer-ho, es generen una sèrie de frases que tenen com a objectiu posar a prova la capacitat dels sistemes de resoldre referències anafòriques per mitjà d'un *challenge test set* (vegeu la secció següent).

4. *Challenge test sets*

Els *test sets* són un conjunt de segments que es fan servir per comprovar la qualitat d'un determinat model o sistema de TA. Aquests conjunts de proves s'han fet servir des dels principis de la TA. Tanmateix, amb l'aparició de la TAN s'han popularitzat, atès que sovint resulta complicat d'entendre exactament com funcionen els algoritmes en la generació del text d'arribada (Ferrando et al., 2022).

Els *challenge test sets* es diferencien perquè, en comptes de presentar una representació més o menys "natural"dels diferents fenòmens que es produeixen en la llengua d'origen, se centren en un fenomen concret (Popovic & Castilho, 2019). Serveixen per estudiar a fons com respon un determinat sistema de TA a un fenomen lingüístic específic. En el nostre cas, la traducció dels pronoms febles. D'aquesta manera, tampoc cal que el nombre de frases que inclou sigui excessivament gran, sinó que en un nombre reduït d'oracions s'incorporen tots els casos complexos per tal de veure si el sistema de TA els resol correctament o quines mancances té.

Els primers *challenge test sets* se centraven en l'estudi de la competència sintàctica dels sistemes de TA basats en regles (Arnold et al., 1993). Amb l'emergència dels sistemes estadístics es va abandonar aquest tipus de sistema per avaluar determinats fenòmens i es va recórrer principalment a les avaluacions automàtiques.

Tanmateix, les millores en la qualitat aportades per la TAN i l'opacitat d'alguns processos en la seva tecnologia van fer sorgir de nou aquest sistema d'avaluació. Alguns d'aquests nous *challenge test sets* avaluen un conjunt de característiques per generar un retrat general del funcionament del sistema de TA (Isabelle et al., 2017). També n'hi ha que analitzen un fenomen general format per diferents subcategories. Sennrich (2017) estudia com els sistemes neuronals modelen fenòmens específics del llenguatge, com la concordança, la producció de noves paraules o la traducció de la polaritat.

Altres *challenge test sets* no es fixen tant en un rendiment global o general del sistema de TA, sinó en la seva capacitat de resoldre una qüestió concreta, com ara l'ambigüitat lèxica dels noms (Rios et al., 2018) o el biaix de gènere (Stanovsky et al., 2019). També n'hi ha que han estudiat la traducció de pronoms. Bawden et al. (2018) avaluen diferents fenòmens discursius com la correferència i la coherència/cohesió lèxica de diferents models neuronals. Guillou & Hardmeier (2016) dissenyen un *challenge test set* per avaluar la traducció de pronoms per diferents sistemes de TA. Conté 250 pronoms i un mètode d'avaluació automàtica que compara la traducció dels pronoms en la sortida de la TA amb una traducció de referència.

Els *challenge tets sets* es poden confeccionar i avaluar de forma manual (Isabelle et al., 2017) o de forma automàtica, tant pel que fa a la creació com a la verificació (Stanovsky et al., 2019). Sovint, però, es barregen els dos mètodes. Des del 2018, a més, s'han inclòs com a tasca d'avaluació a la Conference on Machine Translation[1].

5. Metodologia

Per comprovar com traduïen del castellà al català els pronoms febles sis motors diferents de TA (dos de TABR, dos de TAN i dos GPT) es va optar per crear un *challenge test set* confeccionat *ad hoc* que permetés incorporar les principals dificultats de traducció .[2] L'objectiu és comprovar si els nous models són capaços de millorar la traducció de les anàfores, especialment dels pronoms febles. Perquè els tres sistemes tinguessin els mateixos avantatges, només s'avaluen les anàfores intraoracionals, ja que els sistemes de TABR no tenen en compte cap element fora de la frase.

Es va partir de l'estudi de les combinacions de pronoms febles en català per crear frases *ad hoc* en castellà que recollissin les principals combinacions de pronoms. Es van fer tres grans grups per al conjunt de frases:

Grup 1: Combinacions de dos pronoms febles. Les combinacions binàries de pronoms acostumen a diferir en les dues llengües i això pot plantejar problemes de traducció.

ES Compra el libro a su prima pero no se lo da.

CA Compra el llibre a la seva cosina però no *l'hi* dona.

[1] https://www2.statmt.org/wmt23/testsuite-subtask.html

[2] https://github.com/sergialvarezvidal/test_suite

Grup	Frases	Apertium	Softcatalà	Google	Yandex	ChatGPT	Gemini
1	45	33	27	40	39	33	24
2	45	36	43	24	26	19	23
3	10	4	2	2	0	0	1
Total	**100**	**73**	**72**	**66**	**65**	**52**	**47**
Percentatge		**73%**	**72%**	**66%**	**65%**	**52%**	**47%**

Taula 1: Resultat dels errors per sistema.

Grup 2: Pronoms que no s'expliciten en castellà. Hi ha determinats complements que no queden recollits amb un pronom feble quan es reprenen en castellà, però que resulta obligatori incloure en català. Pot ser que no quedin recollits de cap manera en castellà o que es faci servir el complement explicitat amb un pronom personal o un demostratiu.

ES Le he pedido que se presente al cargo pero no ha accedido.

CA Li he demanat que es presenti al càrrec però no *hi* ha accedit.

ES Quiere mucho a su hija pero no confía en ella.

CA Estima molt la seva filla però no *hi* confia.

Grup 3: Pronoms similars en les dues llengües. Hi ha un gran nombre d'estructures que concorden en les dues llengües i no suposen, a priori, un repte excessiu per a la traducció.

ES Quiere demasiado a Juan y por eso mismo lo odia.

CA Estima massa el Joan i per això mateix l'odia.

Es va confeccionar un conjunt de 100 frases: 45 per al primer grup d'oracions, 45 per al segon i 10 per al tercer. El percentatge de frases atorgat a cada grup té relació amb el nivell de dificultat previst: així, els grups 1 i 2 presenten a priori més divergència entre les dues llengües i es preveu que suposin més problemes de traducció; per tant, obtenen el gruix de frases. El grup 3, en què hi ha frases amb una estructura pronominal similar al castellà, només consta de 10 frases i ens servirà per veure si realment els diferents sistemes resolen fàcilment els pronoms de frases amb estructures similars.

Quant als sistemes de traducció, es van triar sis sistemes, dos per cada model de traducció (TABR, TAN, GPT). Això ens permet avaluar com resolen els diferents models la traducció anafòrica de pronoms i, al mateix, temps, ens permet valorar si hi ha diferències entre les diferents implementacions dels tres models. Es va optar pels models d'ús més habitual, accessibles des de la web. En tots els casos es va fer servir la versió gratuïta. Com que aquests sistemes s'actualitzen tot sovint, tots es van provar el mateix dia, el 2 de febrer de 2024:

- TABR: Apertium[3] i Softcatalà[4]
- TAN: Google Translate[5] i Yandex[6]
- GPT: ChatGPT 3.5[7] i Gemini [8]

6. Resultats

Un cop traduït el conjunt de frases amb els diferents sistemes, es van avaluar manualment els resultats, que es mostren a la Taula 1.

Com es pot veure, els sistemes que proporcionen millors resultats són els models GPT, concretament Gemini, que comet 47 errors (47% del total), seguit de la TAN (66% i 65% d'errors) i la TABR (73% i 72%). Tot i així, a pesar dels bons resultats obtinguts per a aquesta combinació lingüística en les avaluacions automàtiques i manuals (Alvarez et al., 2019), el percentatge d'errors en la resolució de pronoms febles és força alt.

Quant als grups de pronoms febles en què es poden detectar els errors, tots els sistemes tradueixen sense gaires problemes els pronoms en les frases amb una estructura semblant entre el castellà i el català (Grup 3), excepte Apertium. Pel que fa al Grup 1, presenta força problemes per a tots els sistemes. Són el tipus de frases (combinacions de dos pronoms febles) en les quals gairebé tots els sistemes fan més errors, excepte pels sistemes de TABR. Per a aquest grup, Softcatalà és dels sistemes que millor resol la combinació de pronoms, per davant de tots els sistemes de TAN. Per al Grup 2, tot i que també presenta força errors en tots els sistemes, millora respecte al Grup 1. Els diferents sistemes introdueixen

[3] https://www.apertium.org/
[4] https://www.softcatala.org/traductor/
[5] https://translate.google.com/
[6] https://translate.yandex.com/
[7] https://chat.openai.com/
[8] https://gemini.google.com/

1	ES	Los amigos le han pedido la cámara tantas veces, que al final se la ha dejado.
	Correcció	Els amics li han demanat la càmera tantes vegades que al final **els l'**ha deixat.
TABR	Apertium	Els amics li han demanat la càmera tantes vegades que al final la hi ha deixat.
	Softcatalà	Els amics li han demanat la càmera tantes vegades que al final la hi ha deixat.
TAN	Google	Els amics li han demanat la càmera tantes vegades, que al final l'hi ha deixat.
	Yandex	Els amics li han demanat la cambra tantes vegades, que al final la hi ha deixat.
GPT	ChatGPT	Els amics li han demanat la càmera tantes vegades que al final se la ha deixat.
	Gemini	Els amics li han demanat la càmera tantes vegades que al final se la ha deixat.
2	ES	Quiere mucho a su hija pero no confía en ella.
	Correcció	Estima molt la seva filla, però no **hi** confia.
TABR	Apertium	Estima molt la seva filla, però no confia en ella.
	Softcatalà	Vol molt la seva filla però no confia en ella.
TAN	Google	Vol molt la seva filla però no hi confia.
	Yandex	Estima molt la seva filla però no confia en ella.
GPT	ChatGPT	Estima molt la seva filla però no confia en ella.
	Gemini	L'estima molt, però no confia en ella.
3	ES	Salió de casa solo pero entró con María.
	Correcció	Va sortir de casa sol però **hi** va entrar amb la Maria.
TABR	Apertium	Li agradava el teatre, però no hi anava sovint.
	Softcatalà	Va sortir de casa sol però va entrar amb María.
TAN	Google	Va sortir de casa sol però va entrar amb la Maria.
	Yandex	Va sortir de casa sol però va entrar amb La Maria.
GPT	ChatGPT	Va sortir de casa sol però va entrar amb la Maria.
	Gemini	Va sortir de casa sol, però va entrar amb Maria.

Taula 2: Exemples d'errors per a tots els sistemes.

l'ús d'un pronom feble encara que en castellà el complement es repeteix, s'ha elidit o es reprodueix amb l'ús d'un demostratiu.

Els tres sistemes de TA mostren grans problemes per resoldre les frases en les quals cal fer dues substitucions pronominals, que són diferents al castellà (Grup 1). Com es pot veure en el primer exemple de la Taula 2, cap dels tres sistemes no és capaç de fer la substitució correcta de CD i CI (*al final els l'ha deixat*).

En aquests casos, els sistemes acostumen a produir frases calcades del castellà i intenten incloure sovint els pronoms *se* i *ho* con a solucions predeterminades simulant els recursos del castellà. Després d'analitzar els errors d'aquest grup, no hi ha cap complement concret (partitiu de CD, CI plural de tercera persona plural) que produeixi uns resultats que divergeixin dels resultats per a tot el grup.

Pel que fa a les oracions en les quals el castellà no recull el pronom perquè no és necessari però cal explicitar-lo en català (grup 2), el sistema que pitjor resultats produeix és el de TABR. A la traducció al català és necessari recollir amb el pronom *hi* o *en* determinats complements, com el complement d'anar, però com es mostra a l'exemple 2 cap d'aquests sistemes pot resoldre la frase correctament. Per a aquest grup, i seguint amb els calcs detectats com a solucions per al Grup 1, sovint es resolen les traduccions ometent el pronom o utilitzant un demostratiu o pronom personal, com es pot veure en l'exemple 3. Aquesta solució (*confia en ell*) no és pròpia del català i mostra la incapacitat dels sistemes per produir la versió genuïna.

Per obtenir informació sobre els tipus d'errors que cometen els sistemes a l'hora de traduir les oracions amb pronoms febles, s'inclouen dues anàlisis addicionals. D'una banda s'han classificat els errors en omissions (el pronom no hi és o en falta un dels dos necessaris), substitucions (hi ha el nombre necessari de pronoms però són incorrectes) i insercions (s'han afegit pronoms). Com es pot veure a la Taula 3, tot i que hi ha divergències en els percentatges, més de la meitat dels errors per a tots els sistemes tenen a veure amb omissions. Sovint en les combinacions de diversos pronoms, els sistemes només n'inclouen un i en cap cas no afegeixen més pronoms dels necessaris.

Tipus error	Apertium	Softcatalà	Google	Yandex	ChatGPT	Gemini
Omissions	57,5	65,3	51,5	80	55,8	78,7
Substitucions	42,5	34,7	48,5	20	44,2	21,3
Insercions	0	0	0	0	0	0

Taula 3: Tipus d'error expressat en percentatge.

Pronom	Apertium	Softcatalà	Google	Yandex	ChatGPT	Gemini
En	33	35	32	28	36	38
Hi	25	23	38	32	31	35
Ho	0	0	0		0	0
El/la/els/les	34	37	24	28	21	17
Li	8	5	4	16	12	10

Taula 4: Pronoms erronis expressats en percentatge.

També hem anotat quin era el pronom amb el qual es cometia l'error per veure si hi ha una tendència a ometre o usar de forma incorrecte alguns pronoms específics. Com es pot veure a la taula 4, hi ha bastants problemes quan cal incloure els pronoms *en* i *hi*, en molts casos perquè s'ometen i no apareixen en la traducció. L'abundància d'errors en els pronoms determinats està sovint vinculada a les equivocacions que cometen els sistemes de TAN a l'hora de fer la combinació correcta de dos pronoms. Cap dels sistemes analitzats no comet errors a l'hora de col·locar el pronom *ho*. Al contrari, aquest és el pronom que s'inclou per defecte en moltes de les solucions errònies.

7. Conclusions

L'objectiu d'aquest estudi era veure com traduïen els elements anafòrics sis sistemes de traducció automàtica, dos de TABR, dos de TAN i dos models GPT, després de l'èxit que han tingut els models neuronals i els sistemes basats en models massius de llenguatge (MML) tant en les avaluacions automàtiques com manuals, especialment per resoldre problemes de traducció relacionats amb la cohesió textual, com ara l'anàfora, tot i que no han estat dissenyats com a traductors.

Per fer-ho vam confeccionar un *challenge test set*, que és un conjunt de frases que estan dissenyades especialment per posar a prova la capacitat que tenen els sistemes per traduir un fenomen concret. Tot i que aquest conjunt de prova no permet obtenir una avaluació de la qualitat general del sistema, ens pot ajudar a veure com resol un ventall de casos per a un problema concret. A més, el *challenge test set* és públic (així com els resultats obtinguts per a aquests sistemes) i es pot ampliar o es pot provar amb nous motors que es desenvolupin per a aquesta combinació lingüística.

En el cas de les anàfores centrades en els pronoms febles, els resultats dels sis sistemes són decebedors. Tots sis tenen molts problemes per resoldre correctament la traducció de pronoms febles del castellà al català a pesar que els motors de traducció entre aquestes dues llengües obtenen uns altíssims resultats en l'avaluació manual i automàtica.

Dels sis sistemes avaluats, el que més bons resultats dona és el model GPT, concretament Gemini. Això confirmaria la recerca recent sobre les millores d'aquest model en la traducció d'elements cohesius del text Castilho et al. (2023). Tot i així, el millor sistema falla en gairebé la meitat de les oracions. Les propostes errònies mostren una traducció que tendeix a ser força literal del castellà, bé perquè fa servir els pronoms de la frase original o perquè inclou una estructura sintàctica calcada del castellà que omet l'ús dels pronoms febles.

Aquests resultats no ens permeten treure conclusions sobre el funcionament general dels sistemes de traducció automàtica avaluats, però ens permeten veure com es comporten davant d'un problema complex de traducció. Una qüestió que cal tenir en compte, més enllà del nombre d'errors, és la influència excessiva de l'estructura sintàctica del castellà en les propostes de traducció al català.

Aquest fenomen és precisament el que pensem abordar en la nostra futura recerca, és a dir, veure si l'estructura del text original influeix en excés en les propostes de traducció dels diferents models en la combinación del castellà al català.

Agraïments

Aquest treball ha rebut suport parcial del projecte TAN-IBE: Traducció automàtica neuronal per a les llengües romàniques de la Península Ibèrica finançat pel Ministerio de Ciencia e Innovación. Proyectos de generación de conocimiento 2021. Referència: PID2021-124663OB-I00.

Referències

Alvarez, Sergi, Antoni Oliver & Toni Badia. 2019. Does NMT make a difference when post-editing closely related languages? The case of Spanish-Catalan. En *Machine Translation Summit XVII: Translator, Project and User Tracks*, 49–56.

Aranberri, Nora, Gorka Labaka, Arantza Díaz de Ilarraza & Kepa Sarasola. 2017. Ebaluatoia: crowd evaluation for English–Basque machine translation. *Language Resources and Evaluation* 51(4). 1053–1084. 10.1007/s10579-016-9335-x

Arnold, Doug, Dave Moffat, Louisa Sadler & Andrew Way. 1993. Automatic test suite generation. *Machine Translation* 8(1/2). 29–38.

Bahdanau, Dzmitry, Kyunghyun Cho & Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv [cs.CL]* 10.48550/arXiv.1409.0473

Barrault, Loïc, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post & Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (wmt). En *5th Conference on Machine Translation*, 1–55.

Bawden, Rachel, Rico Sennrich, Alexandra Birch & Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. En *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1304–1313. 10.18653/v1/N18-1118

Bayatli, Sevilay, Sefer Kurnaz, Ilnar Salimzianov, Jonathan North Washington & Francis M. Tyers. 2018. Rule-based machine translation from Kazakh to Turkish. En *21st Annual Conference of the European Association for Machine Translation*, 49–58.

Bentivogli, Luisa, Arianna Bisazza, Mauro Cettolo & Marcello Federico. 2018. Neural versus phrase-based MT quality: An in-depth analysis on English–German and English–French. *Computer Speech & Language* 49. 52–70. 10.1016/j.csl.2017.11.004

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever & Dario Amodei. 2020. Language models are few-shot learners. *arXiv [cs.CL]* 10.48550/arXiv.2005.14165

Brussel, Laura, Arda Tezcan & Lieve Macken. 2018. A fine-grained error analysis of NMT, PBMT and RBMT output for English-to-Dutch. En *11th International Conference on Language Resources and Evaluation (LREC)*, 3799–3804.

Buysschaert, Joost, María Fernández-Parra, Koen Kerremans, Maarit Koponen & GysWalt van Egdom. 2018. L'acceptació de la disrupció digital en la formació en traducció: la immersió tecnològica en simulacions de despatxos de traducció. *Tradumàtica tecnologies de la traducció* 16. 125–133. 10.5565/rev/tradumatica.209

Castilho, Sheila, Clodagh Quinn Mallon, Rahel Meister & Shengya Yue. 2023. Do online machine translation systems care for context? what about a GPT model? En *24th Annual Conference of the European Association for Machine Translation*, 393–417.

Castilho, Sheila, Joss Moorkens, Federico Gaspari, Rico Sennrich, Vilelmini Sosoni, Yota Georgakopoulou, Pintu Lohar, Andy Way, Antonio Valerio, Antonio Valerio Miceli Barone & Maria Gialama. 2017. A comparative quality evaluation of PBSMT and NMT using professional translators. En *Machine Translation Summit XVI*, 116–131.

Costa-jussà, Marta R. 2017. Why Catalan–Spanish neural machine translation? analysis, comparison and combination with standard rule and phrase-based technologies. En *4th Workshop on NLP for Similar Languages, Varieties and Dialects*, 55–62. 10.18653/v1/W17-1207

van Deemter, Kees & Rodger Kibble. 2000. On coreferring: Coreference in MUC and related annotation schemes. *Computational Linguistics* 26(4). 629–637.

Do Campo Bayón, María & Pilar Sánchez-Gijón. 2019. Evaluating machine translation in a low-resource language combination: Spanish–Galician. En *Machine Translation Summit XVII*, 30–35.

España-Bonet, Cristina, Gorka Labaka, Arantza Díaz de Ilarraza & Lluís Màrquez. 2011. Hybrid machine translation guided by a rule–based system. En *Machine Translation Summit XIII*,

Ferrando, Javier, Gerard I. Gállego, Belen Alastruey, Carlos Escolano & Marta R. Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. En *Conference on Empirical Methods in Natural Language Processing*, 8756–8769. 10.18653/v1/2022.emnlp-main.599

Fité Labaila, Ricard. 2001. La traducció automàtica aplicada a la premsa escrita. El cas d'El Periódico en català. *Treballs de Comunicació* 21–25.

Fité Labaila, Ricard. 2007. Cas d'integració de la TA: El Periódico. *Tradumàtica: traducció i tecnologies de la informació i la comunicació* 4.

Forcada, Mikel L. 2017. Making sense of neural machine translation. *Translation Spaces* 6(2). 291–309. 10.1075/ts.6.2.06for

Forcada, Mikel L., Mireia Ginestí-Rosell, Jacob Nordfalk, Jim O'Regan, Sergio Ortiz-Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez & Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation* 25. 127–144. 10.1007/s10590-011-9090-0

Ge, Niyu, John Hale & Eugene Charniak. 1998. A statistical approach to anaphora resolution. En *6th Workshop on Very Large Corpora*, 161–170.

Guillou, Liane & Christian Hardmeier. 2016. PROTEST: A test suite for evaluating pronouns in machine translation. En *10th International Conference on Language Resources and Evaluation (LREC)*, 636–643.

Hendy, Amr, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify & Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *arXiv [cs.CL]* 10.48550/arXiv.2302.09210

Hobbs, Jerry R. 1978. Resolving pronoun references. *Lingua* 44(4). 311–338. 10.1016/0024-3841(78)90006-2

Institut d'Estudis Catalans. 2016. *Gramàtica de la llengua catalana*. Institut d'Estudis Catalans

Isabelle, Pierre, Colin Cherry & George Foster. 2017. A challenge set approach to evaluating machine translation. En *Conference on Empirical Methods in Natural Language Processing*, 2486–2496. 10.18653/v1/D17-1263

Islam, Md. Adnanul, Md. Saidul Hoque Anik & A. B. M. Alim Al Islam. 2022. An enhanced RBMT: When RBMT outperforms modern data–driven translators. *IETE Technical Review* 39(6). 1473–1484. 10.1080/02564602.2022.2026828

Kehler, Andrew, Douglas Appelt, Lara Taylor & Aleksandr Simma. 2004. The (non)utility of predicate-argument frequencies for pronoun interpretation. En *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, 289–296.

Koponen, Maarit, Leena Salmi & Markku Nikulin. 2019. A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Machine Translation* 33. 61–90. 10.1007/s10590-019-09228-7

Lappin, Shalom & Herbert J. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics* 20(4). 535–561.

Mitkov, Ruslan. 1999. Introduction: Special issue on anaphora resolution in machine translation and multilingual NLP. *Machine Translation* 14(3). 159–161. 10.1023/A:1011132522992

Mitkov, Ruslan, Sung-Kwon Choi & Randall Sharp. 1995. Anaphora resolution in machine translation. En *6th Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*, 87–95.

Popovic, Maja & Sheila Castilho. 2019. Challenge test sets for MT evaluation. En *Machine Translation Summit XVII*, presentation.

Rios, Annette, Mathias Müller & Rico Sennrich. 2018. The word sense disambiguation test suite at WMT18. En *3rd Conference on Machine Translation*, 588–596. 10.18653/v1/W18-6437

Sennrich, Rico. 2017. How grammatical is character–level neural machine translation? Assessing MT quality with contrastive translation pairs. En *15th Conference of the European Chapter of the Association for Computational Linguistics*, 376–382.

Sghaier, Mohamed Ali & Mounir Zrigui. 2020. Rule-based machine translation from Tunisian dialect to modern standard Arabic. *Procedia Computer Science* 176. 310–319. 10.1016/j.procs.2020.08.033

Stanovsky, Gabriel, Noah A. Smith & Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. En *57th Annual Meeting of the Association for Computational Linguistics*, 1679–1684. 10.18653/v1/P19-1164

Tognini-Bonelli, Elena. 2001. *Corpus linguistics at work*. John Benjamins Publishing Company. 10.1075/scl.6

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. *arXiv [cs.CL]* 10.48550/arXiv.1706.03762

Voita, Elena, Pavel Serdyukov, Rico Sennrich & Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. En *56th Annual Meeting of the Association for Computational Linguistics*, 1264–1274. 10.18653/v1/P18-1117

Wicks, Rachel & Matt Post. 2022. Does sentence segmentation matter for machine translation? En *7th Conference on Machine Translation*, 843–854.

Novas Perspectivas

Propuesta recibida en febrero 2024 y aceptada para publicación en mayo 2024.

Explorando las capacidades de los modelos de lenguaje neuronales en la identificación y clasificación de colocaciones léxicas

Exploring the capabilities of neural language models for the identification and classification of lexical collocations

Radovan Milović
Universidad de Santiago de Compostela

Resumen

La mayoría de las investigaciones sobre el procesamiento automatizado de colocaciones se ha centrado en el uso de medidas de asociación. Sin embargo, el enfoque se ha ido cambiando lentamente hacia la exploración de la efectividad de los modelos de lenguaje neuronales o *neural language models* (NLMs). En este artículo, investigamos el último método mediante el ajuste fino de modelos de la familia BERT en inglés, español y portugués utilizando recursos léxicos anotados con Funciones Léxicas (FL). Examinamos así las capacidades de los modelos de lenguaje para la identificación y clasificación de colocaciones léxicas tanto en escenarios monolingües como multilingües. Los resultados de los desempeños generales variaron, con valores F que oscilan entre 0.30 y 0.51. Concluimos que el modelo multilingüe sobresale en el aprendizaje cruzado al emplear un conjunto de entrenamiento combinado de los tres idiomas. Además, a pesar de la posible variabilidad, los resultados demuestran una mejor identificación de las Funciones Léxicas con un mayor número de instancias en el conjunto de entrenamiento. Por último, realizamos un análisis cualitativo para investigar posibles patrones de identificación errónea exhibidos por el modelo.

Palabras clave

colocaciónes léxicas, funciones léxicas, modelos de lenguaje neuronales, ajuste fino

Abstract

The majority of research on automated collocation processing has focused on using association measures. However, the focus has been slowly shifting to exploring the effectiveness of neural language models (NLMs). In this paper, we investigate the latter by fine-tuning BERT family models in English, Spanish, and Portuguese using annotated lexical resources with Lexical Functions (LFs). We examine the capabilities of language models for the identification and classification of lexical collocation in both monolingual and multilingual scenarios. The results of the overall performances varied, with f1 scores ranging from 0.30 to 0.51. We conclude that the multilingual model excels in cross-lingual learning by employing a combined training set of all three languages. Moreover, despite possible variability, the results demonstrate improved identification of Lexical Functions with a larger number of instances in the training set. Lastly, we conduct a qualitative analysis to investigate possible patterns of misidentification exhibited by the model.

Keywords

lexical collocations, lexical functions, neural language models, fine-tuning

1. Introducción

Según Pawley (1985, p. 102), el lenguaje debe considerarse "una colección de formas de hablar sobre las cosas [...], expresando ideas de una manera convencional (gramatical, idiomática, etc.)"[1], destacando la noción de que la estructura del lenguaje se construye a través de patrones recurrentes que surgen de su uso. Estos patrones, influenciados por las convenciones de un idioma específico, exhiben una arbitrariedad inherente. A nivel léxico, esto es particularmente evidente en combinaciones de palabras como *lluvia torrencial*, *dormir profundamente*, *prestar atención* y otras similares. Estas expresiones son conocidas por su complejidad lingüística, ya que el significado de los constituyentes de las frases "torrencial", "profundamente" y "prestar" se desvía de sus definiciones literales cuando se encuentran en contacto con una palabra específica. Además, la variación en la expresión de conceptos similares a través de diferentes idiomas, como *prestar atención* en español, *Aufmerksamkeit schenken* en alemán o *pay attention* en inglés,

[1] Cita original: "[...] a collection of ways of talking about things [...], expressing ideas in a manner that is conventional (grammatical, idiomatic, etc.)" (Pawley, 1985)

DOI: 10.21814/lm.16.1.428

los hace impredecibles. Tales frases idiosincráticas son comúnmente conocidas como "colocaciones."

La adquisición de colocaciones representa un desafío significativo específicamente para los estudiantes de segundo idioma, ya que requiere un adecuado dominio de los matices lingüísticos específicos de cada idioma. Para superar este desafío, los aprendices suelen recurrir a diccionarios. Sin embargo, antes de incorporar información sobre colocaciones en los diccionarios, es necesario elaborar listas de colocaciones extrayéndolas de corpus. Por lo tanto, los desafíos planteados por las colocaciones también se extienden al ámbito del procesamiento del lenguaje natural (PLN). Además del procesamiento computacional de colocaciones con fines lexicográficos, estas son un factor crucial en una variedad de tareas de PLN como el aprendizaje de idiomas asistido por ordenador, la traducción automática y la desambiguación del sentido de las palabras.

La investigación en PLN sobre colocaciones gira predominantemente en torno al empleo de diversas medidas de asociación para extraer listas de candidatos a colocaciones (Church & Hanks, 1990; Smadja, 1993; Rychlý, 2008), que posteriormente son evaluadas. Este enfoque se ha convertido en el método principal para el procesamiento automático de colocaciones. Sin embargo, descuida las características semánticas fundamentales de las colocaciones. Las medidas de asociación no logran discriminar las colocaciones de otras expresiones multipalabra, lo que lleva a una lista de candidatos que abarca una variedad de coocurrencias.

Investigaciones más recientes han comenzado a explorar estrategias para el descubrimiento de colocaciones que se centran en el uso de *word embeddings* (Rodríguez-Fernández et al., 2016; Enikeeva & Mitrofanova, 2017) y los modelos de lenguaje neuronales (Espinosa-Anke et al., 2021, 2022; Nisho, 2022). Los *word embbedings* y los modelos de lenguaje neuronales mapean la similitud semántica en un espacio vectorial multidimensional según sus patrones de distribución en los corpus, permitiendo así considerar las propiedades semánticas. Además, estos estudios hacen uso de recursos léxicos que contienen colocaciones anotadas con FL. Como resultado, la tarea de detectar colocaciones en corpus con respecto a sus propiedades semánticas se ha entrelazado con su categorización. El objetivo de este enfoque, que también es el punto focal de nuestra investigación, es identificar instancias de colocaciones en corpus y clasificarlas simultáneamente según el modelo teórico de las Funciones Léxicas.

En este artículo, llevamos a cabo un experimento exploratorio que implica el ajuste fino de modelos de lenguaje neuronales con recursos léxicos anotados con FL en tres idiomas: inglés, español y portugués. Estos corpus fueron creados por García et al. (2019), especialmente para abordar el área poco explorada del procesamiento multilingüe de colocaciones. Por consiguiente, queremos tener en cuenta tanto configuraciones monolingües como multilingües, lo que nos permite obtener percepciones sobre el rendimiento de estos modelos en diferentes contextos lingüísticos. Observamos que los datos multilingües generalmente mejoran la identificación de colocaciones en la lengua meta. Sin embargo, algunas FL parecen más difíciles de aprender que otras, debido a la cantidad de instancias de esa FL a las que el modelo está expuesto durante el entrenamiento. A través del análisis cualitativo, también intentamos descubrir conocimientos lingüísticos mediante la identificación de patrones de error presentes en las predicciones del modelo.

2. Marco teórico

2.1. La noción de "colocación"

La definición de la noción de "colocación" no tiene consenso, principalmente debido a diferencias en las perspectivas y métodos de investigación. Sin embargo, el discurso académico actual reconoce dos interpretaciones principales: la interpretación estadística y la interpretación semántica.

Desde el punto de vista estadístico, las colocaciones se perciben como un fenómeno empírico que nos permite aprender sobre los patrones de comportamiento de una palabra (el nodo) en relación con las palabras adyacentes (los colocativos) (Evert, 2004). Estas colocaciones estadísticas se definen como coocurrencias que aparecen en textos más de lo esperado por azar (Sinclair, 1991). La noción de "colocación estadística" sirve como principio fundamental para su extracción automatizada mediante la aplicación de medidas de asociación.

Nuestro trabajo, sin embargo, gira en torno a la interpretación semántica de las colocaciones, las cuales se definen en función de la composición estructural de sus componentes. Comúnmente se les denomina "colocaciones léxicas" (Krenn, 2000) para distinguirlas de la definición estadística. Desde este punto de vista, las colocaciones son construcciones binarias léxicamente vinculadas, que consisten en una **base** y un **colocativo**, cuya relación muestra disparidad en términos

de aspectos sintácticos y semánticos (Mel'čuk, 1996; Hausmann, 1998). La base se elige libremente según la intención del hablante de transmitir un significado particular, mientras que la elección del colocado está restringida por las convenciones del lenguaje.

El principio fundamental de este enfoque es definir las colocaciones diferenciándolas de las combinaciones libres y las locuciones, principalmente en función de los niveles de composicionalidad y sustituibilidad (Nesselhauf, 2005).

Mientras que una combinación libre no tiene restricciones y opacidad semántica, una locucion, por ejemplo *dar en el clavo*, carece de transparencia semántica, ya que su significado "hacer o decir algo correctamente"no puede determinarse solo analizando sus componentes, lo que resulta en una completa falta de composicionalidad. Además, sus componentes no pueden ser sustituidos por elementos sinónimos para transmitir el mismo significado. Por otro lado, una colocación como *dar un paseo* muestra una composicionalidad parcial[2], ya que uno de sus constituyentes ("paseo") lleva un significado semántico transparente, mientras que la interpretación del otro constituyente ("dar") sigue siendo más ambigua. La elección del colocativo también está influenciada por la convención de un idioma particular y no puede ser sustituida sin alterar el significado de toda la frase. En consecuencia, las colocaciones están limitadas a una combinación interdependiente específica y se caracterizan por una transparencia semántica parcial y una falta de libertad combinatoria.

2.2. Clasificación de colocaciones

La clasificación de colocaciones suele basarse en sus patrones sintácticos, según la categoría gramatical de sus constituyentes: *lluvia torrencial* (adjetivo + sustantivo), *prestar atención* (verbo + sustantivo), *dormir profundamente* (verbo + adverbio), etc. (Hausmann, 1998, p. 1010; Benson et al., 2010). Sin embargo, existe otro sistema de clasificación más exhaustivo y detallado desarrollado por Igor Mel'čuk (1996), llamado Funciones Léxicas (FL).

[2]Las perspectivas sobre la noción de composicionalidad también pueden variar de un autor a otro. Por ejemplo, Mel'čuk (2012, p. 39) considera las colocaciones como frases enteramente composicionales, ya que su significado puede dividirse en dos partes de modo que correspondan a los dos constituyentes. Sin embargo, la comprensión general de este enfoque es que las colocaciones representan el área intermedia de un continuo de transparencia, con combinaciones libres y locuciones ubicadas en los dos extremos (Dražić, 2014, p. 17).

Las Funciones Léxicas proporcionan una clasificación concisa de las genuinas relaciones léxicas, que pueden ser tanto sintagmáticas (relaciones de colación) como paradigmáticas (relaciones semánticas). En el contexto de este trabajo, nuestro enfoque se centra en las FL sintagmáticas, es decir, las colocaciones léxicas.

Para su representación sistemática, se ilustran diferentes tipos de relaciones como una notación matemática $\mathbf{f(L) = Li}$. En términos de FL sintagmáticos, el símbolo **f** representa una función léxica particular, **L** denota el constituyente elegido libremente o *keyword* (base), y **Li** denota los elementos restringidos o *values* (colocativos).

El objetivo de esta clasificación es capturar el significado subyacente inferido por los colocativos y su relación de dependencia de la base. Por ejemplo, la FL **Magn** representa un grupo de colocativos que comparten el mismo significado cuando se asocian a una base. Así, en inglés, *deeply, heartily* y *terribly* son colocativos de *sorry* que comparten el mismo significado subyacente de "intensidad" representado como **Magn**(sorry) = {deeply, heartily, terribly}, capturando simultáneamente la semántica de la colocación y la interdependencia entre los constituyentes. De la misma manera, FL **Bon** (p. ej., *diálogo fructífero*) expresa "bueno" o FL **Oper1** (p. ej., *prestar atención*) "hacer". Además, aunque la elección de los colocativos puede diferir entre idiomas, puesto que expresan el mismo significado central, las Funciones Léxicas pueden aplicarse universalmente a todos los idiomas:

$$\mathbf{Oper1}(\text{atención}) = \{\textit{prestar}/\textit{pay}/\textit{schenken}\}\,.$$

Este modelo teórico ha encontrado relevancia en el procesamiento del lenguaje natural debido a sus diversas propiedades (Kolesnikova, 2011, p. 68–71): su encapsulación de las propiedades sintácticas y semánticas de las colocaciones las hace valiosas para resolver ambigüedades sintácticas y léxicas, mientras que su aplicabilidad universal las convierte en una herramienta ideal para la traducción automática. Además, la clasificación detallada de construcciones lingüísticas impredecibles las hace útiles para crear programas de aprendizaje de idiomas que podrían ayudar a los estudiantes a adquirir tales estructuras. Por último, se vuelven centrales para la identificación y clasificación automáticas en corpus, lo cual exploramos más a fondo en este trabajo.

3. Trabajos relacionados

La identificación de colocaciones en relación con las medidas de asociación ha ganado la mayor popularidad hasta el momento. Esta línea de investigación progresó desde la información mutua (Church & Hanks, 1990), hasta diversas medidas de asociación utilizadas hoy en día, como *log-Dice*, *log-likelihood*, *t-score* y otras (Evert, 2004; Rychlý, 2008). Además, estas medidas se han complementado con etiquetador morfosintáctico (Evert & Kermes, 2003), analizadores de dependencias (Lin, 1999; Seretan & Wehrli, 2006), y finalmente con medidas direccionales (Gries, 2013; Carlini et al., 2014). A través de esta metodología, los candidatos a colocaciones pasan por un riguroso proceso de filtrado, abordando también los problemas de discontinuidad y asimetría de los constituyentes colocacionales.

En cuanto a la tarea autónoma de clasificar automáticamente las colocaciones basadas en el modelo de Funciones Léxicas, los métodos tempranos utilizaron representaciones semánticas basadas en hiperónimos, como *WordNet*, en conjunto con técnicas de aprendizaje automático (Wanner, 2004; Wanner et al., 2006; Gelbukh & Kolesnikova, 2012). Sin embargo, con los avances en la representación semántica y la introducción de *word embeddings* (modelo *Word2vec* de Mikolov et al. (2013)), la tarea de identificar y clasificar colocaciones se unificó.

Uno de los primeros estudios en clasificar e identificar colocaciones simultáneamente fue realizado por Rodríguez-Fernández et al. (2016). Combinan *Word2vec* y el modelo teórico de Funciones Léxicas para identificar y clasificar colocaciones. El método del estudio consiste en utilizar el algoritmo *Word2vec* para generar *word embbedings* en las que las relaciones entre las bases (por ejemplo, "thought" en *deep thought* o "wind" en *strong wind*) y las glosas ("intensity") de sus colocativos correspondientes se mapean en consecuencia. El objetivo es utilizar esta información semántica capturada por los *word embbedings* para recuperar los colocativos potenciales dado una nueva base y glosa.

Los avances en PLN alcanzaron su punto máximo con el desarrollo de modelos de lenguaje neural utilizando la arquitectura de transformador (Vaswani et al., 2017). Lo que hace que estos modelos sean de última generación es su capacidad para enfocarse atentamente en diferentes segmentos de la secuencia de entrada y, por lo tanto, tener en cuenta todo el contexto. Espinosa-Anke et al. (2021) llevaron a cabo dos experimentos para evaluar la eficacia de los transformadores en el manejo de colocaciones. El primer experimento, realizado en una configuración no supervisada, implicó enmascarar el colocativo dentro de una colocación dada y evaluar la precisión predictiva del modelo de lenguaje. El experimento subsiguiente, realizado en una configuración supervisada, se centró en el ajuste fino de los modelos para predecir la función léxica asociada con las colocaciones.

A continuación, Espinosa-Anke et al. (2022) mejoraron el rendimiento de los tranformadores al tener en cuenta las relaciones de dependencia entre la base y el colocativo, integrando un *Graph-aware Transformer* (transformador sensible al grafo) en la estructura del modelo, diseñado específicamente para el análisis de dependencias. Además, incorporan un clasificador de oraciones para determinar la presencia de colocaciones en las oraciones, ofreciendo un contexto adicional para su identificación. Finalmente, utilizan grandes corpus anotados con funciones léxicas en inglés, español y francés para entrenar el modelo y evaluar la efectividad de la arquitectura modificada en la identificación y clasificación de colocaciones.

Junto a los estudios mencionados, es importante destacar la investigación que explora la utilización de *word embeddings* para la tarea respectiva en el idioma ruso realizada por Enikeeva & Mitrofanova (2017), así como la aplicación del *Graph-aware Transformer* para el procesamiento de colocaciones japonesas por Nisho (2022).

Los estudios descritos han mostrado niveles variables de éxito. Además, la falta de transparencia del funcionamiento interno de estos modelos presenta una barrera significativa para los investigadores que buscan comprender precisamente cómo operan estos modelos y por qué producen ciertas salidas. Dicho esto, la identificación y clasificación de colocaciones léxicas utilizando NLMs sigue siendo un territorio abierto que requiere una mayor exploración.

4. Experimento

El ajuste fino es un método popular en el aprendizaje automático, mediante el cual un modelo se entrena en conjuntos de datos más pequeños para adaptarlo específicamente a una tarea objetivo. En esta sección, proporcionamos una descripción detallada de los pasos tomados para el ajuste fino de modelos de lenguaje neural para aprender patrones de colocaciones.

	EN			ES			PT		
	train	**dev**	**test**	**train**	**dev**	**test**	**train**	**dev**	**test**
#tokens	41483	4993	5121	30316	3794	3796	46082	5711	5676
#orac. sin FL	2276	458	255	956	112	114	1710	248	288
#orac. con FL	342	29	51	232	29	25	391	49	45

Cuadro 1: Estadísticas del conjunto de datos para cada idioma.

4.1. Conjunto de datos

Para los experimentos, utilizamos corpus anotados con funciones léxicas en inglés, español y portugués (García et al., 2019)[3], que constan de más de 155.000 tokens y 1.526 colocaciones clasificadas en 60 funciones léxicas. Con el fin de ajuste fino para la clasificación de tokens, convertimos los corpus en un sistema de etiquetado inspirado en el formato BIO. En este formato, el token que denota el elemento base de la colocación se anotó como "B-{FL}", mientras que el token que representa el colocativo se marcó como "C-{FL}". Todos los demás tokens dentro de una oración se etiquetaron como "O" ("outside") para indicar que no pertenecen a la colocación.

Token	**Etiqueta**
Harry	O
felt	O
a	O
hot	C-Magn
surge	B-Magn
of	O
anger	O
.	O

Cuadro 2: Oración de ejemplo con nuevas etiquetas.

Para evitar posibles confusiones, se eliminaron las colocaciones anidadas (p. ej., una expresión en inglés "take a deep breath", donde tanto "take (a) breath" como "deep breath" son colocaciones) y múltiples instancias de las mismas FL dentro de la misma oración (un total de 11 oraciones). Dejamos para futuros trabajos una exploración de estos casos, utilizando, por ejemplo, el análisis sintáctico de dependencia para identificar las relaciones base-colocativo. El Cuadro 1 muestra las estadísticas del conjunto de datos final para cada idioma, mientras que la Figura 1 presenta la frecuencia relativa de las funciones léxicas más comunes.

[3]Para obtener la clasificación completa de FL utilizada en los corpus, así como la descripción del enfoque de anotación y los enlaces a los corpus, consulte a García et al. (2019).

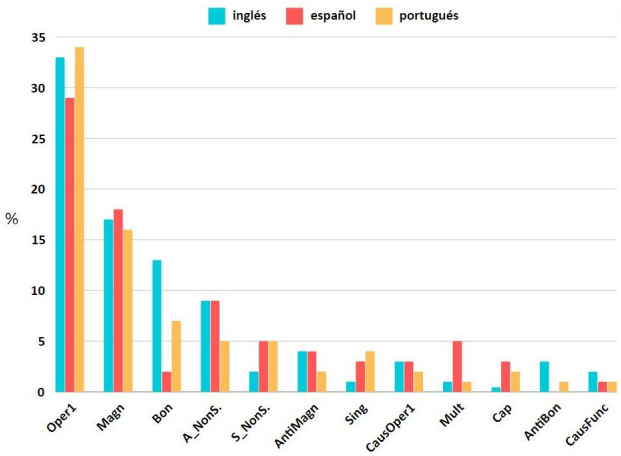

Figura 1: Distribución de las funciones léxicas más comunes para cada idioma.

Dividimos el conjunto de datos en subconjuntos de entrenamiento, validación y prueba según el número total de tokens para cada idioma. El propósito del conjunto de entrenamiento es permitir que el modelo aprenda y capture patrones y características que ayudarán en la identificación de colocaciones. Durante la fase de entrenamiento, el modelo utiliza el conjunto de validación para evaluación interna, refinando continuamente sus parámetros a lo largo de múltiples *epochs* (ciclos de aprendizaje). Una vez que el modelo ha seleccionado la mejor configuración de parámetros, se evalúa el rendimiento en datos de prueba desconocidos para evaluar sus capacidades de generalización.

El 80 % de los tokens se utilizaron para el conjunto de entrenamiento, seguido por el conjunto de validación con el 10 % subsiguiente y el 10 % restante para el conjunto de prueba. Además, nos aseguramos de que el conjunto de entrenamiento abarcara todas las funciones léxicas que aparecieran en los restantes subconjuntos. Los conjuntos de entrenamiento consistían en instancias de colocaciones únicas que no aparecían en los conjuntos de validación y prueba, asegurando así que los modelos no memorizaran simplemente colocaciones específicas, sino que aprendieran los patrones lingüísticos subyacentes.

4.2. Modelos

Para nuestra tarea, utilizamos una familia de modelos BERT y sus *base*[4] variantes adaptadas para diferentes idiomas:

- **BERT** (Devlin et al., 2018) para inglés;
- **RoBERTa** (Gutiérrez-Fandiño et al., 2022) para español;
- **BERTimbau** (Souza et al., 2020) para portugués;
- **mBERT** (Devlin et al., 2018) para la configuracion multilingüe.

Además, para comparar el rendimiento de los modelos de transformadores evaluados, entrenamos el modelo **BiLSTM-CNN-CRF** (Chernodub et al., 2019)[5] con los mismos datos. Este modelo se entrena durante un mínimo de 50 *epochs*.

4.3. Ajuste fino

Para el proceso de ajuste fino empleamos el script *run_ner*[6] (Wolf et al., 2020), diseñado específicamente para el ajuste fino de tareas de clasificación de tokens. En nuestro experimento, utilizamos los parámetros predeterminados, que incluyeron ejecutar el proceso de ajuste fino durante 3 *epochs*.

4.4. Evaluación

Para la evaluación utilizamos el script *conlleval*[7]. El código adopta como entrada las etiquetas verdaderas y predichas y calcula métricas de evaluación, la precisión, la exhaustividad y el valor F. Genera un resumen general del rendimiento y el rendimiento para cada tipo de entidad, que en nuestro caso se corresponde con las funciones léxicas.

4.5. Configuraciones experimentales

Con el propósito de explorar las capacidades de generalización de los modelos en reconocimiento de colocación, llevamos a cabo diferentes configuraciones experimentales. En primer lugar, realizamos el ajuste fino de modelos monolingües utilizando datos etiquetados para cada idioma. Después, reproducimos este proceso con un modelo multilingüe. Finalmente, diseñamos un enfoque de aprendizaje cruzado para entrenar un modelo multilingüe con datos combinados de diferentes idiomas. Por lo tanto, el experimento involucró las siguientes configuraciones:

- Ajuste fino de modelos monolingües **BERT**, **RoBERTa** y **BERTimbau** para cada idioma.
- Ajuste fino del modelo multilingüe mBERT con datos monolingües: **mBERT-mono**.
- Ajuste fino de mBERT utilizando conjuntos de entrenamiento combinados de inglés, español y portugués: **mBERT-multi**.

5. Resultados y análisis

El Cuadro 3 ofrece una visión general de los resultados de rendimiento, mostrando valores F en un rango de 0.30 a 0.51.

5.1. Análisis del rendimiento general

5.1.1. Modelos monolingües

El modelo de referencia, que utiliza la arquitectura BiLSTM-CNN-CRF, tuvo en rendimiento deficiente (promedio $F = 0,11$). Por el contrario, los modelos de transformadores ajustados superaron significativamente el modelo de referencia, lo que indica que la arquitectura de transformadores del modelo entrenado proporcionó beneficios en términos de reconocimiento de colocación. Teniendo en cuenta los parámetros de entrenamiento, la familia de modelos BERT se entrenó durante solo tres *epochs*, mientras que el modelo de referencia requirió alrededor de 30 *epochs* para lograr su máximo rendimiento en los tres lenguajes. Esta notable superioridad de los modelos BERT sobre el modelo de referencia enfatiza las capacidades significativamente mayores de los modelos transformadores para comprender señales contextuales y su adaptabilidad para capturar diversas estructuras lingüísticas.

También es importante destacar que el modelo RoBERTa mostró un rendimiento inferior ($F = 0,32$) en comparación con los modelos BERT inglés ($F = 0,45$) y portugués ($F = 0,47$). La diferencia principal entre los modelos BERT y RoBERTa es el tamaño de los datos de preentrenamiento. RoBERTa se entrena en un corpus de preentrenamiento más grande, lo que potencialmente puede proporcionar una representación lingüística más rica. Sin embargo, Pérez-Mayos et al. (2021) exploraron recientemente la

[4] La variante *base* se refiere a la configuración estándar de un modelo.

[5] https://github.com/achernodub/targer

[6] https://github.com/huggingface/transformers/tree/main/examples/pytorch/token-classification

[7] https://github.com/sighsmile/conlleval

	EN			ES			PT		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
(Ro)BERT(a/imbau)	0.69	0.34	0.45	0.44	0.25	0.32	0.56	0.40	**0.47**
mBERT-mono	0.51	0.27	0.35	0.46	0.23	0.30	0.41	0.37	0.39
mBERT-multi	0.69	0.41	**0.51**	0.48	0.44	**0.46**	0.48	0.42	0.45
BiLSTM-CNN-CRF	0.13	0.13	0.13	0.23	0.07	0.11	0.14	0.07	0.09

Cuadro 3: Rendimiento general de las configuraciones para cada idioma.

correlación entre el tamaño de los datos de preentrenamiento y el rendimiento de los modelos de lenguaje neuronales en la adquisición de patrones sintácticos. Concluyeron que "si bien los modelos preentrenados con más datos codifican más conocimientos sintácticos y tienen un mejor rendimiento en aplicaciones posteriores, no siempre ofrecen un mejor rendimiento en diferentes fenómenos sintácticos". Por lo tanto, el impacto del tamaño del corpus de preentrenamiento en el rendimiento del modelo puede variar según la tarea objetivo. Las puntuaciones más bajas de los modelos RoBERTa en comparación con los modelos BERT pueden sugerir que más datos de preentrenamiento no siempre conducen a mejores resultados para la identificación de colocaciones.

5.1.2. Modelos multilingües

La comparación entre los modelos monolingües y mBERT-mono revela que los modelos monolingües muestran un rendimiento superior en comparación con mBERT cuando se entrenan con datos monolingües. Esto concuerda con hallazgos anteriores (Singh & Lefever, 2022; Conneau et al., 2022) que indican que la abrumadora cantidad de idiomas en los datos de preentrenamiento de modelos multilingües puede conducir a la dilución de información en tareas monolingües. Sin embargo, dada su capacidad para codificar información multilingüe, parecen intrigantes para evaluar enfoques de aprendizaje cruzado, como se demostró en el experimento final (mBERT-multi).

La configuración mBERT-multi, en donde realizamos el ajuste fino de mBERT con conjuntos de entrenamiento combinados, muestra un aumento significativo de rendimiento en la mayoría de los aspectos (excepto la precisión de BERTimbau). Esta mejora podría atribuirse a dos factores clave. En primer lugar, la inclusión de conjuntos de entrenamiento adicionales lleva a un aumento sustancial en el tamaño de los datos. El conjunto de datos ampliado proporciona a mBERT una representación más amplia de patrones lingüísticos, mejorando su capacidad para generalizar en varios idiomas. En segundo lugar, a diferencia de la configuración mBERT-mono, donde llevamos a cabo el ajuste fino del modelo con datos monolingües, la configuración mBERT-multi expone al modelo a una variedad de idiomas, lo que le permite participar en el aprendizaje entre idiomas, capturando así mejor las propiedades sintáctico-semánticas de las colocaciones presentadas por Funciones Léxicas universalmente aplicables.

5.2. Análisis a nivel de FL

Para realizar un análisis a nivel de FL, nos centraremos en una configuración específica, mBERT-multi, que logró predominantemente los mejores resultados generales.

El Cuadro 4 muestra el valor F para cada FL. En algunos casos (Oper1 y Magn), los resultados son altos (0.60 o más), mientras que para varias FL, los modelos no pudieron reconocer correctamente ni una sola colocación. Para explorar más a fondo esta discrepancia, realizamos un análisis de correlación entre la frecuencia de las FL en los datos de entrenamiento y su rendimiento.

5.2.1. Análisis de correlación

Utilizando la correlación de *Spearman*, investigamos la correlación entre el número de tipos de colocaciones en los datos de entrenamiento y el rendimiento de los modelos en el reconocimiento de las respectivas funciones léxicas. Con este análisis, pretendemos determinar si el número de ejemplos de FL se correlaciona con el rendimiento del modelo en el reconocimiento de FL, o si el rendimiento se atribuye a las capacidades del modelo para generalizar las características lingüísticas intrínsecas de FL.

El coeficiente de correlación de 0.70 y 0.72 para inglés y español (Cuadro 4) indica una correlación positiva fuerte entre las FL. Esto sugiere que una mayor frecuencia de FL en los datos de entrenamiento corresponden con un mejor rendimiento en el reconocimiento de las FL respectivas para ambos idiomas. Por lo tanto, aumentar el número de ejemplos de FL en los datos de entrena-

		EN	ES	PT
	cuenta	**valor F**		
Oper1	352	0.60	**0.68**	0.60
Magn	189	**0.72**	0.60	**0.69**
Bon	88	0.56	-	0.44
A_NonS.	87	0.38	0.00	0.00
S_NonS.	47	0.00	0.25	0.00
AntiMagn	40	0.00	0.66	-
CausOper1	33	0.00	0.00	0.00
Sing	31	0.00	0.00	0.00
Cap	25	0.00	0.00	0.00
Mult	22	0.00	0.00	0.00
AntiBon	19	0.28	-	-
CausFunc	19	-	0.00	0.66
correlación		0.70	0.72	0.37
valor p		0.02	0.02	0.29

Cuadro 4: Cuentas y rendimientos de FL más frecuentes en la configuración mBERT-multi, resultados de la correlación de *Spearman* y valor p.

miento probablemente mejorará la competencia del modelo en el reconocimiento de FL. A pesar de que las correlaciones no son estadísticamente significativas ($p = 0{,}2$, probablemente debido al pequeño número de ejemplos), los valores de *Spearman* son notablemente altos. Sin embargo, el coeficiente de correlación de 0.37 en portugués revela una correlación positiva más débil entre la frecuencia de colocación de FL y el rendimiento del modelo. Como podemos ver, a pesar de que el conjunto de entrenamiento contenía un bajo número de ejemplos de CausFunc, el modelo pudo identificar sus instancias, dando como resultado un valor F elevado de 0.66. Aunque aún hay una asociación positiva, no es tan fuerte como se observa en inglés y español. Esto sugiere que, aunque un aumento en la frecuencia de colocación de FL en los datos de entrenamiento puede tener un efecto positivo en el rendimiento del modelo, otros factores como la capacidad de los modelos para aprender las propiedades sintáctico-semánticas de FL pueden contribuir a su identificabilidad.

5.2.2. Análisis de errores

El modelo mostró un rendimiento más alto al identificar la función léxica Magn, logrando un valor F promedio de 0.67 en los tres idiomas. FL Oper1 fue la segunda función léxica mejor identificada, con un promedio de valor F de 0.63. En los párrafos siguientes, nuestra atención se centrará en estas FL. Dado que ocurren con mayor frecuencia, nos proporcionarán información suficiente para el análisis de errores. Buscamos identificar posibles inclinaciones de aprendizaje erróneas del modelo. Las oraciones a continuación ejemplifican los principales desafíos encontrados por los modelos:[8]

(1) China has been a [**great** $_{C-Bon;\ C-Magn}$] [**help** $_{B-Bon;\ B-Magn}$] at getting North Korea to the table...[9]

Problemas para distinguir entre Magn y otras funciones léxicas semánticamente similares, como Bon. A medida que el modelo aprendía más sobre otras funciones léxicas, más dificultades tenía para distinguirlas. Las funciones léxicas Magn y Bon están intrincadamente conectadas y describen típicamente colocaciones que involucran patrones de "adjetivo + sustantivo". Estas funciones buscan capturar el significado fundamental transmitido por colocativos o modificadores que expresan diferentes modificaciones. Magn se refiere principalmente a modificaciones cuantitativas, mientras que Bon está asociada con calificaciones subjetivas. Debido a su similitud, los modelos a veces tienen dificultades para distinguir entre estas dos funciones léxicas. Por ejemplo, en casos como *great help*, los modelos reconocen erróneamente esta colocación como Magn. Dado que "great" funciona como un adjetivo que denota algo de gran magnitud, los modelos no logran discriminarlo de Bon, donde "great" se usa como un modificador positivo cualitativo. Esta confusión puede surgir de las diferencias poco claras en los significados semánticos centrales de los colocativos y la similitud en el patrón sintáctico entre estas dos funciones léxicas.

(2) ...mandam os casos mais graves para um [**grande** $_{O;\ C-Magn}$] [**depósito** $_{O;\ B-Magn}$], que conhecemos como presídio. [10]

Problemas para identificar colocativos con cambio semántico vago. El modelo tuvo éxito en general al reconocer las propiedades semánticas centrales de esta clase de colocaciones. Sin embargo, hubo algunas instancias de falsos positivos, como *grande depósito*. Podemos ver que el modelo tiene dificultades para distinguir estos adjetivos cuando se usan como parte de combinaciones libres versus como

[8] La primera etiqueta es la etiqueta original y la segunda es la etiqueta predicha por el modelo.

[9] Traducción: "China ha sido de gran ayuda para llevar a Corea del Norte a la mesa de negociaciones..."

[10] Traducción: "...mandamos los casos más graves a un gran depósito, que conocemos como presidio."

constituyentes en la colocación Magn. Una posible razón para esta confusión es que el cambio de significado entre su uso independiente y su papel como constituyentes de la colocación no es lo suficientemente pronunciado. Por ejemplo, tanto "grande" en una combinación libre como "grande" como parte de la colocación Magn transmiten la noción de algo de gran magnitud, aunque con matices sutiles. Otro posible factor es la frecuencia de ocurrencia de estos colocativos. Dado el uso frecuente de estos tipos de adjetivos como intensificadores en la colocación, el modelo puede aprender a asociarlos más fuertemente con esta función léxica, lo que podría llevar a falsos positivos.

(3) . . . que están realmente [**enfrentando** $_{C-Oper1;\ O}$] [**problemas** $_{B-Oper1;B-Oper1}$] por la pobreza", indicó.

Problemas para identificar colocativos de Oper1. Los colocativos de Oper1, comúnmente conocidas como verbos de apoyo, según Mel'čuk (1996, p. 53), se consideran "semánticamente vacíos". Aunque el cambio semántico en este caso es muy evidente, podemos ver la incertidumbre del modelo al identificar específicamente los colocativos.

(4) Assim como pode impulsionar as vendas de uma empresa, uma celebridade pode [**causar** $_{C-CausFunc1;\ C-CausOper1}$] o [**efeito** $_{B-CausFunc1;\ B-Oper1}$] contrário.[11]

Problemas para distinguir entre Oper1 y otras FL que representan verbos de apoyo, como Func. A diferencia de Magn y Bon, que se distinguen por criterios semánticos, la diferenciación entre Oper1 y otras funciones léxicas de verbos de apoyo se basa en patrones sintácticos. La función léxica Caus ("iniciar una situación") se combina a menudo con otras funciones léxicas para formar expresiones complejas de FL. Aunque el modelo demostró un mejor aprendizaje de las implicaciones semánticas de la función léxica Caus, este éxito llevó a una confusión adicional al diferenciar entre Oper1, CausFunc y CausOper1.

(5) . . . General Chen Zaido del EPL (militar chino) decidió [**dar** $_{O;\ C-Oper1}$] [**espalda** $_{O;\ B-Oper1}$] en la facción de Guardas Rojas moderado Millón Heroes[. . .]

Problemas para diferenciar entre colocaciones y locuciones. Algunas instancias presentan desafíos para determinar si una frase debe categorizarse como una colocación o una locucion. Para el modelo, un ejemplo de dicho caso es la expresión en español *dar (la) espalda*. Esta frase se caracteriza por una total idiomaticidad, ya que su significado de "rechazar" no se puede inferir a partir de sus partes constituyentes, siendo por lo tanto considerada una locucion. Sin embargo, el patrón estructural de la frase se asemeja al patrón sintáctico de una función léxica Oper1, lo que causa confusión en el modelo. Como resultado, el modelo la identifica como Oper1, teniendo en cuenta el patrón sintáctico de la colocación e incluso reconociendo su naturaleza idiomática, pero sin identificarla correctamente.

Una de las características destacadas de los modelos de transformadores es su capacidad para considerar todo el contexto oracional en lugar de observar únicamente palabras vecinas. Teniendo eso en cuenta, también es necesario señalar la posibilidad de que los patrones de aprendizaje del modelo estén influenciados por las señales contextuales que rodean a las colocaciones. Ya sea Oper1 o Magn, o cualquier otra función léxica, estas señales contextuales desempeñan un papel significativo en la formación de la comprensión e interpretación del modelo sobre las colocaciones. Sin embargo, debido a la diversidad de oraciones en los corpus, no está claro cómo afecta exactamente el contexto a la identificación precisa de funciones léxicas.

6. Conclusiones

El reconocimiento automático de colocaciones léxicas, particularmente con una pequeña cantidad de datos de entrenamiento, es una tarea difícil. Las colocaciones son expresiones lingüísticas complejas y existen innumerables combinaciones únicas a considerar. Sin embargo, podemos ver que los modelos con arquitectura transformadora son capaces de aprender mejor que la generacion de modelos anterior.

Los enfoques de aprendizaje cruzado generalmente mejoran el rendimiento de los modelos multilingües, lo que indica su potencial para capturar propiedades semántico-sintácticas de colocaciones entre idiomas. Se observa también un mayor éxito en el reconocimiento de diferentes tipos de colocaciones con un mayor número de instancias de una función léxica particular en los datos de entrenamiento, aunque con excepciones.

[11] Traducción: "Así como puede impulsar las ventas de una empresa, una celebridad puede causar el efecto contrario."

Con los rápidos avances en el aprendizaje automático, la investigación futura aún espera explorar nuevas arquitecturas y combinaciones de datos de entrenamiento, teniendo en cuenta todos los desafíos potenciales que los modelos puedan enfrentar. Además, investigar los mecanismos internos de los NLMs nos acercará a mejorar y comprender las capacidades de los modelos de lenguaje neuronal en la identificación y clasificación de colocaciones.

Agradecimientos

Este artículo surgió a partir de TFM del Máster Europeo en Lexicografía. Quiero expresar mi agradecimiento a todos los afiliados del programa por su orientación y asistencia a lo largo de todo el proceso.

Referencias

Benson, Morton, Evelyn Benson & Robert Ilson. 2010. *The BBI combinatory dictionary of English.* John Benjamins Publishing Company. 10.1075/z.bbi

Carlini, Roberto, Joan Codina-Filba & Leo Wanner. 2014. Improving collocation correction by ranking suggestions using linguistic knowledge. En *3rd Workshop on NLP for Computer-Assisted Language Learning*, 1–12.

Chernodub, Artem, Oleksiy Oliynyk, Philipp Heidenreich, Alexander Bondarenko, Matthias Hagen, Chris Biemann & Alexander Panchenko. 2019. TARGER: Neural argument mining at your fingertips. En *57th Annual Meeting of the Association for Computational Linguistics*, 195–200. 10.18653/v1/P19-3031

Church, Kenneth Ward & Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics* 16(1). 22–29

Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer & Veselin Stoyanov. 2022. Unsupervised cross-lingual representation learning at scale. En *58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451. 10.18653/v1/2020.acl-main.747

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. En *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186. 10.18653/v1/N19-1423

Dražić, Jasmina. 2014. *Leksičke i gramatičke kolokacije u srpskom jeziku.* Filozofski fakultet Novi Sad.

Enikeeva, Ekaterina & Olga Mitrofanova. 2017. Russian collocation extraction based on word embeddings. En *International Conference Dialogue 2017: Computational Linguistics and Intellectual Technologies*, 52–64.

Espinosa-Anke, Luis, Joan Codina-Filba & Leo Wanner. 2021. Evaluating language models for the retrieval and categorization of lexical collocations. En *16th Conference of the European Chapter of the Association for Computational Linguistics*, 140–104. 10.18653/v1/2021.eacl-main.120

Espinosa-Anke, Luis, Alexander Shvets, Alireza Mohammadshahi & Leo Wanner. 2022. Multilingual extraction and categorization of lexical collocations with graph-aware transformers. En *11th Joint Conference on Lexical and Computational Semantics*, 89–100. 10.18653/v1/2022.starsem-1.8

Evert, Stefan. 2004. *The statistics of word cooccurrences: Word pairs and collocations*: University of Stuttgart. Tesis Doctoral

Evert, Stefan & Hannah Kermes. 2003. Experiments on candidate data for collocation extraction. En *10th Conference of The European Chapter of the Association for Computational Linguistics*, 83–86

García, Marcos, Marcos García-Salido, Susana Sotelo, Estela Mosqueira & Margarita Alonso-Ramos. 2019. Pay attention when you pay the bills. a multilingual corpus with dependency-based and semantic annotation of collocations. En *57th Annual Meeting of the Association for Computational Linguistics*, 4012–4019. 10.18653/v1/P19-1392

Gelbukh, Alexander & Olga Kolesnikova. 2012. *Semantic analysis of verbal collocations with lexical functions.* Springer. 10.1007/978-3-642-28771-8

Gries, Stefan Th. 2013. 50-something years of work on collocations: What is or should be next.... *International Journal of Corpus Linguistics* 18(1). 137–166. 10.1075/ijcl.18.1.09gri

Gutiérrez-Fandiño, Asier, Jordi Armengol-Estape, Marc Pamies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre & Marta Villegas. 2022. MarIA: Spanish language models. *Procesamiento Del Lenguaje Natural* 68. 39–60. 10.26342/2022-68-3

Hausmann, Franz Josef. 1998. Le dictionnaire de collocations. En *Wörterbücher, Dictionaries, Dictionnaires. Ein internationales Handbuch zur Lexikographie*, 1010–1019. Walter de Gruyter

Kolesnikova, Olga. 2011. *Automatic extraction of lexical functions*: Instituto Politecnico Nacional — Centro de Investigacion en Computacion. Tesis Doctoral.

Krenn, Brigitte. 2000. CDB: A database of lexical collocations. En *2nd International Conference on Language Resources and Evaluation*,

Lin, Dekang. 1999. Automatic identification of noncompositional phrases. En *37th Annual of the Association for Computational Linguistics (ACL)*, 317–324. 10.3115/1034678.1034730

Mel'čuk, Igor. 1996. Lexical functions: A tool for the description of lexical relations in a lexicon. En *Lexical Functions in Lexicography and Natural Language Processing*, 37–102. John Benjamins

Mel'čuk, Igor. 2012. Phraseology in the language, in the dictionary, and in the computer. *Yearbook of Phraseology* 3(1). 31–56. 10.1515/phras-2012-0003

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado & Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. En *Advances in Neural Information Processing Systems*, 3111–3119

Nesselhauf, Nadja. 2005. *Collocations in a learner corpus*. John Benjamins. 10.1075/scl.14

Nisho, Kosuke James. 2022. *Extraction and categorization of Japanese lexical collocations with graph-aware transformers*: Universidad Pompeu Fabra. Trabajo de Fin de Máster

Pawley, Andrew. 1985. On speech formulas and linguistic competence. *Lenguas Modernas* 12. 84–104

Pérez-Mayos, Laura, Miguel Ballesteros & Leo Wanner. 2021. How much pretraining data do language models need to learn syntax? En *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 7927–7934. 10.18653/v1/2021.emnlp-main.118

Rodríguez-Fernández, Sara, Luis Espinosa-Anke, Roberto Carlini & Leo Wanner. 2016. Semantics-driven recognition of collocations using word embeddings. En *54th Annual Meeting of the Association for Computational Linguistics (ACL)*, 499–505. 10.18653/v1/P16-2081

Rychlý, Pavel. 2008. A lexicographer-friendly association score. En *Recent Advances in Slavonic Natural Language Processing*, 6–9.

Seretan, Violeta & Eric Wehrli. 2006. Accurate collocation extraction using a multilingual parser. En *21st International Conference on Computational Linguistics and 44th Annual of the Association for Computational Linguistics*, 317–324. 10.3115/1220175.1220295

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford University Press

Singh, Pranaydeep & Els Lefever. 2022. When the student becomes the master: Learning better and smaller monolingual models from mBERT. En *29th International Conference on Computational Linguistics*, 4434–4441.

Smadja, Frank. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics* 19(1). 143–178.

Souza, Fabio, Rodrigo Nogueira & Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. En *Brazilian Conference on Intelligent Systems*, 403–417. 10.1007/978-3-030-61377-8_28

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aaidan N. Gomez, Lukasz Kaiser & Illia Polosukhin. 2017. Attention is all you need. En *31st Conference on Neural Information Processing Systems*, 6000–6010. 10.5555/3295222.3295349

Wanner, Leo. 2004. Towards automatic fine-grained semantic classification of verb-noun collocations. *Natural Language Engineering* 10(2). 95–143. 10.1017/S1351324904003328

Wanner, Leo, Bernd Bohnet & Mark Giereth. 2006. Making sense of collocations. *Computer Speech & Language* 20(4). 609–624. 10.1016/j.csl.2005.10.002

Wolf, Tomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf,

Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest & Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. En *Conference on Empirical Methods in Natural Language Processing*, 38–45. 10.18653/v1/2020.emnlp-demos.6

http://www.linguamatica.com/

linguaMATICA