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.