Projecte GenAI

CatZh – Traducció per LLMs especializat en corpus de salut

***Pipeline* per a la modificació de la base de dades**

S’entén la pipeline com a un recorregut que cada unitat semàntica (vectors) passarà automàticament per tal de modificar la base de dades català-xinès mandarí simplificat (CAT>ZH) resultant del projecte AINA del Barcelona SuperComputer (BSC). Aquesta modificació pretén ésser una millora quant a la precisió de la traducció del vocabulari i, a més a més, una base per a la seva ampliació.

En aquest document s’entendrà el català com a CAT, el castellà espanyol peninsular com a ES, l’anglès americà com a ENG i el xinès mandarí simplificat com a ZH. El lemes extrets de la base de dades glossari tindran un subíndex 1 i els lemes resultants de la pipeline presentaran un subíndex 2. Així doncs, el mot extret d’AINA en CAT serà CAT1 i el mot resultant del seu processament per la pipeline serà CAT2, en cas que el resultat fos en CAT.

Existeixen diferents opcions de recorregut de la pipeline:

* **Opció 1:**

ZH1 🡪 ENG 🡪 ZH2

Comparativa entre ZH1 y ZH2 a favor del ZH2.

* **Opció 2:**

CAT1 🡪 ENG 🡪 ZH2

* **Opció 3:**

CAT1 🡪 ES 🡪 ENG 🡪 ZH2

* **Opció 4:**

CAT1 🡪 ES 🡪 ZH2

En cas que la varietat de llengües nodreixi el LLM en desenvolupament, es recomana l’opció 3 (Op3), atès que les interferències lingüístiques entre CAT i ES son comunes i, sovint, indetectables per a la majoria dels catalanoparlants del territori. A més a més, la naturalesa de l’àrea de coneixement implica que els coneixements lingüístics en anglès americà també seran útils.

En cas que la varietat de llengües implicades en la pipeline que es proposa no fos rellevant per al nodriment lingüístic de la LLM en qüestió, **es recomana l’opció 4** (Op4). Malgrat que el corpus que nodreix les eines de traducció automàtica (TA) que s’usaran en aquesta pipeline és, en principi, més ric quant a ZH-ENG, l’eina escollida per al pas ES>ZH també té una gran precisió, segons diverses proves que s’han dut a terme. A més a més, les eines que s’encarreguen de les fases CAT>ENG i CAT>ES són molt més efectives en el segon cas (també, vistos els resultats de diverses proves), per tant, es descarta l’Op2 a favor de l’Op4.

Eines proposades:

SoftCatalà (CAT>ES)

<https://www.softcatala.org/traductor/>

DeepL (ES>ZH)

<https://www.deepl.com/es/translator>

Ambdues eines de TA s’han construït a partir de xarxes neuronals i funcionen a partir d’estadístiques quant a l’ús. Per contra, existeix l’eina alternativa al portal SoftCatalà, i que també té l’API exigida per les característiques del projecte, Apertium (<https://www.apertium.org/index.spa.html#?dir=cat-spa&q=>), però s’ha descartat amb motiu de la seva construcció a partir de regles i no, de xarxes neuronals.

Així mateix, es recomana l’ús d’eines de correcció ortotipogràfica sobre el mot CAT1 i també sobre el mot ZH2, resultant de la *pipeline*. En aquest cas, es recomana l’ús de LanguageTool (<https://languagetool.org/ca>), que presenta l’API exigida i corregeix tant CAT com ZH per ús estadístic i a partir d’una xarxa neuronal d’IA.

En resum, el mot inicial hauria de passar per LanguageTool, ajustat a CAT; hauria de traduir-se a ES amb SoftCatalà; es traduiria a ZH amb DeepL, i tornaria a passar per LanguageTool, ajustat a ZH.