Viquitrobada 14 de novembre 2020

Corpus de veu lliures amb Common Voice

Joan Montané

SOFTCATALÀ
www.softcatala.org

Hola!

Sóc el Joan

Membre voluntari de @softcatala

jmontane@softcatala.org

@unjoanqualsevol

Softcatalà

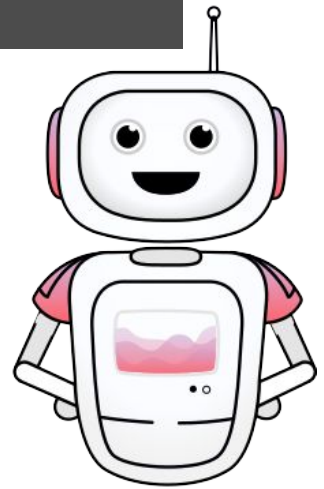
Softcatalà és una associació sense ànim de lucre, que té l'objectiu de fomentar l'ús del català a la informàtica, Internet i les noves tecnologies

- Traducció de programes
- Recursos per a traductors
- Recursos lingüístics
- Altres recursos

Dades de veu lliures: Mozilla Common Voice

Objectiu: creació de corpus de veu lliures (CC0) per a qualsevol llengua de manera comunitària.

- Inici del #CommonVoiceCAT: juliol 2018
- 700 hores enregistrades / 578 h. validades
- 5.200 col·laboradors
- 1r objectiu: 1.000 hores validades
- 2n objectiu: 10.000 hores validades



I després del Common Voice?

Objectiu final: que hi hagi corpus de dades lingüístiques per al català amb llicència lliure, i serveis que els usin

- **Accessibilitat:**

  Síntesi de veu

  Reconeixement automàtic de la parla

- **Mycroft en català**
- **Serveis de veu de Mozilla en català**

#CommonVoiceCAT

Els col·laboradors poden enregistrar o validar talls de veu.

https://commonvoice.mozilla.org/ca

- Fàcil i còmode des de PC, mòbil o tauleta
- Dades anònimes. Sempre.
- Opcionalment: perfil, dades demogràfiques
- Volem totes les variants dialectals
- +18 anys

Com s'hi pot col·laborar?

- Difonent el projecte Common Voice en la comunitat viquipedista
- Enregistrant talls de veu
- Validant talls de veu
- Aportant-hi textos moderns amb llicència CC0
- Incrementant la variació dialectal

Gràcies!

@unjoanqualsevol

SOFTCATALÀ
www.softcatala.org

El traductor neural anglès català lliure de Softcatalà: Funcionament i oportunitats

Viquitrobada 14 de novembre 2020

Jordi Mas, jmas@softcatala.org

Hola!

Sóc en Jordi Mas

Membre voluntari de @ **Softcatalà**

Àrea d'enfocament: processament del llenguatge natural

jmas@softcatala.org

jordimash

Què veurem avui

1. El traductor què i per què
2. Qualitat de les traduccions
3. Oportunitats per a la Viquipèdia

1

El traductor què i per què

Context

Existeixen 3 tecnologies de traducció:

- Sistemes basats en regles (Apertium)
  - Són molt ràpids, ideals per llengües amb pocs recursos (no requereixen corpus), funcionen bé amb llengües properes (entre llengües romàniques)
  - Cal molt d'esforç per definir regles de traducció entre llengües llunyanes
- Sistemes estadístics
  - Aprenen a partir de grans volums d'exemples (corpus)
- Sistemes neuronals
  - Aprenen a partir de grans volums d'exemples (corpus)
  - Són la tecnologia actual usada per Google Translate, Microsoft, i similars

Què estem construint?

- Un traductor neuronal anglès - català
- https://www.softcatala.org/traductor-angles-catala/
  - Podeu traduir textos anglès - català directament
  - També permetem traduir fitxers de text en UTF-8 fins a 256Kb
- Que tota aquesta tecnologia sigui lliure:
  - https://github.com/Softcatala/nmt-softcatala
  - Corpus lliures
  - Tecnologies lliures (OpenNMT, etc)
  - Publicació de models lliures
  - Ho podeu adaptar i executar en la vostra màquina

Per què aquest projecte?

- Aconseguim un millor rendiment en la traducció català - anglès que amb Apertium i oferir-lo als usuaris
- Oferir una tecnologia neuronal lliure i gratuita que els usuaris puguin usar fàcilment
  - No existeix una tecnologia lliure
  - Actualment tots els traductors neuronals cobren per traducció (a través d'API)
- Utilitzar models neurals per a accelerar la traducció de nous projectes de traducció a Softcatalà

2

Qualitat de les traduccions

Avaluació quantitativa

- Tenim una frase en anglès i traduïda al català per un humà
- Després demanem a la màquina que tradueixi l'anglès
- Com que sabem com ha d'estar traduït en català, la qualitat es determina segons la similitud entre la traducció automàtica i la humana

Problemes:

- Una mateixa frase pot estar traduïda correctament de diverses maneres

Avaluació quantitativa

- BLEU (*bilingual evaluation understudy*)
  - BLEU simplement calcula la precisió de n-gram (petit fragment d'una paraula) afegint un pes igual a cadascun
- NIST
  - Mateix principi que BLEU però té en compte com d'informatiu és l'n-gram. Quan es troba un n-gram correcte, com més rar sigui l'n-gram, més pes se li donarà
- Sistemes dissenyats per a ser una mesura de corpus no de frases

Avaluació qualitativa

- Humans valoren la traducció i indiquen si és correcta
- Es pot fer de diverses maneres
  - Demanant a traductors que evaluin el resultat
  - Recollint l'activitat de l'usuari en producció
    - Per exemple, demanant-li valorar la traducció

Situació actual parell anglès - català

	Corpus tatoeba		Corpus SleepyHollow	
	BLEU	**NIST**	**BLEU**	**NIST**
Neuronal de Softcatalà	0.38	7.99	0.15	4.84
Apertium (SC)	0.19	5.24	0.07	3.50
Yandex	0.28	6.30	0.10	3.90
Google Translator	0.33	7.29	0.14	4.61

- És relatiu dir que un sistema és millor que un altre sistema, perquè depèn bastant dels corpus (cas d'ús). Dit això, així es valora en les competicions internacionals.

3

Oportunitats per a la Viquipèdia

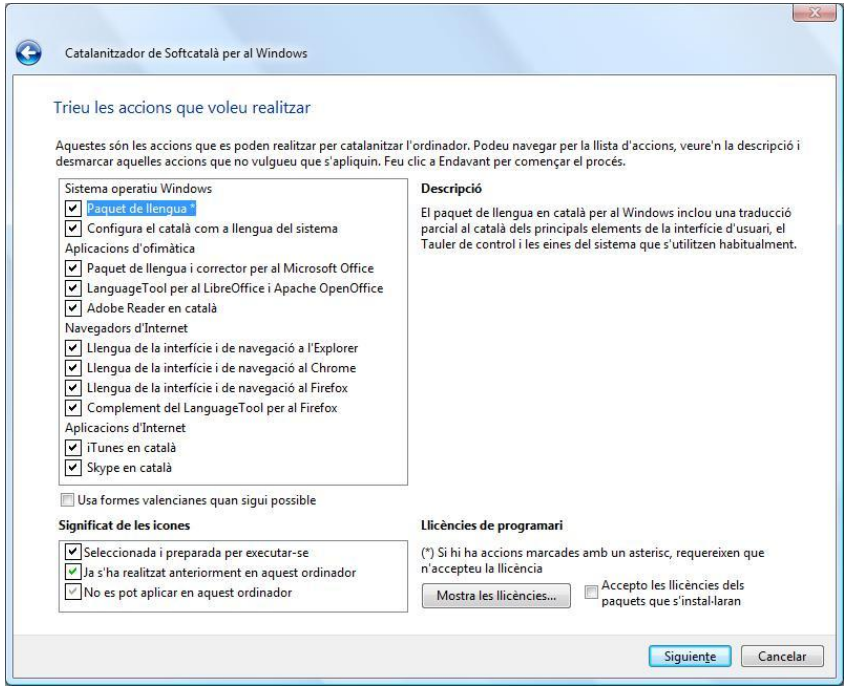
Oportunitats per a Viquipèdia

- Integració amb Content Translator
  - Permetria que aquest motor estigués disponible per a la comuntitat de traductors
- Treballar amb Amical Wikimedia per entrenar models específics per als editors de Viquipèdia
  - Això pot suposar una millor qualitat ja que són models entrenats específicament per a un domini
- Podem entrenar altres parells de llengües
  - P.ex: català - alemany

Oportunitats per a viquipedistes

- Podeu usar la tecnogia a través de la web
  - https://www.softcatala.org/traductor-angles-catala/
- Podeu integrar l'eina amb els vostres fluxos d'edició
  - És compatible amb l'API d'Apertium

Per últim... el Catalanitzador



- Si els usuaris no naveguen en català no arriben a la Viquipèdia en català
- El Catalanitzador configura automàticament totes les aplicacions, navegadors i diccionaris en català
- Disponible per al Windows i macOS: https://www.softcatala.org/catalanitzador/
- En mòbils, les guies que expliquen com configurar-los pas a pas https://www.softcatala.org/ordinadors-i-mobils-en-catala/tutorials/

Gràcies!

Comentaris? Preguntes?

jordimash