Traducción automática neuronal con MTUOC

Antoni Oliver - Universitat Oberta de Catalunya (UOC)

08/04/2021

Webinar Alumni

Índice de la presentación

Alumni

Introducción

Objetivos de la presentación

- Repasar las diferentes estrategias de T.A.
- Presentar las ventajas de TAN frente a otras estrategias de T.A.
- ¿Cómo podemos utilizar T.A.N. en nuestros proyectos?
- Sistemas comerciales
- Toolkits de T.A.N.
- El proyecto MTUOC
- Creación de un sistema T.A.N. paso a paso
- Otros proyectos similares
- Conclusiones

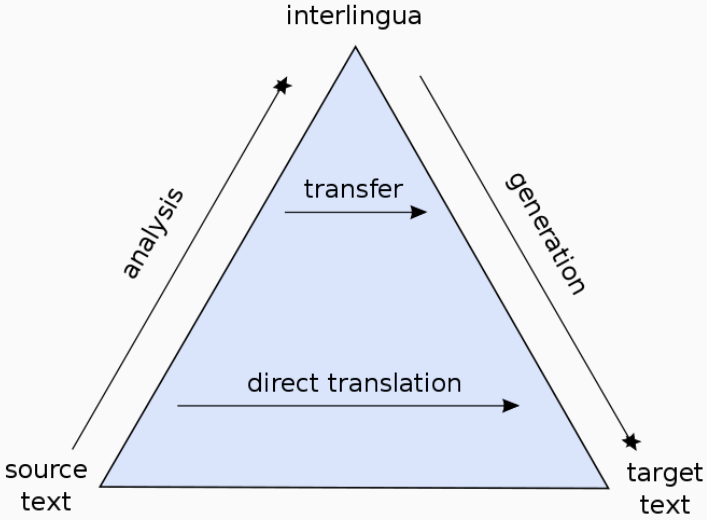
UOC Alumni

Las estrategias de TA

Estrategias de TA

- Basadas en reglas
  - Traducción directa
  - Transferencia
  - Interlingua
- Basadas en corpus
  - Estadística
  - Neuronal

Estrategias basadas en reglas



Traducción directa

O: This is a simple test.

T: this is a simple test .

Diccionario:

this: esto

is: es

simple: sencillo

example: prueba

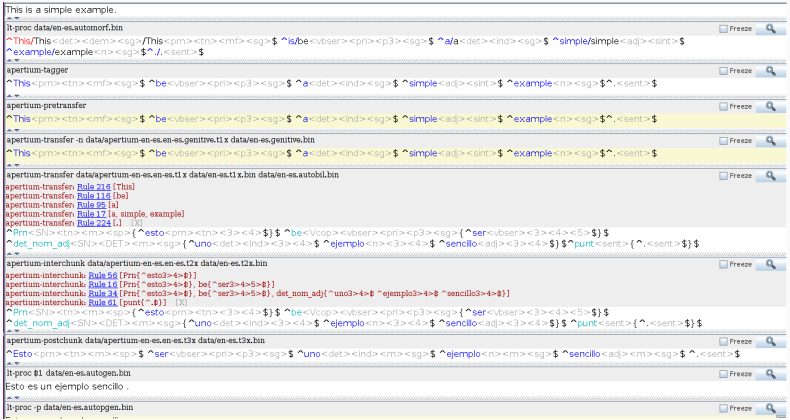
TD: esto es un sencillo prueba

Reglas para concordancia

Reglas para cambio de orden

TD2: esto es una prueba sencilla

Transferencia sintáctica



Traducción Estadística

this is a ||| se trata de un ||| 0.317073 0.0194084...

this is a ||| es este un ||| 0.428571

a simple example ||| un ejemplo sencillo ||| 1 0.139799...

this is ||| esto es ||| 1 ...

...

simple example ||| ejemplo sencillo ||| 0.8 0.247418 ...

...

example ||| ejemplo ||| 0.594592 ...

Traducción neuronal

Se basa en una representación vectorial (word embeddings) de las palabras. P.E. la palabra ”example” se puede representar por un vector (en el ejemplo 300 dimensiones):

[2.05078125e-01 7.85827637e-04 3.54003906e-02 1.00585938e-01 -5.44433594e-02 1.53320312e-01 2.55859375e-01 -2.18750000e-01 -3.31115723e-03 2.09960938e-01 -2.07031250e-01 1.77001953e-02 4.29687500e-02 -2.01171875e-01 -1.57226562e-01 1.88476562e-01 ...]

Estos vectores en cierta manera representan el significado de las palabras.

Ejemplo: palabras similares

Las 10 palabras más similares a ”example” son:

- (’instance’, 0.7873880863189697)
- (’examples’, 0.604103684425354)
- (’illustration’, 0.5342041254043579)
- (’exemplar’, 0.4963938295841217)
- (’shining_example’, 0.489044189453125)
- (’reason’, 0.4628060758113861)
- (’Example’, 0.44629594683647156)
- (’counterexample’, 0.4366326332092285)
- (’analogy’, 0.43353956937789917)
- (’metaphor’, 0.4081893861293793)]

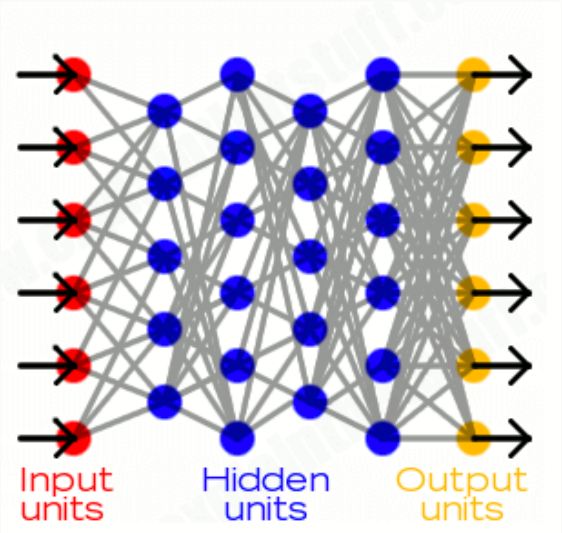
Ejemplo: operaciones con vectores

woman + king - man

Se obtiene:

[('queen', 0.7118193507194519)]

Red neuronal



Comparativa eng-spa

This is a simple example.

Apertium eng-spa (transferencia sintáctica superficial)

Esto es un ejemplo sencillo .

Moses (estadístico, entrenado con un corpus de 5 M de segmentos)

Este es un ejemplo sencillo.

Marian (neuronal, entrenado con un corpus de 18 M de segmentos)

Este es un ejemplo sencillo.

Comparativa eng-spa

Autism is a lifelong neurological condition that manifests during early childhood, irrespective of gender, race, or socio-economic status.

Apertium eng-spa (transferencia sintáctica superficial)

El autismo es un lifelong neurological condición que manifiesta durante niñez temprana, irrespective de género, carrera, o socio-estado económico.

Moses (estadístico, entrenado con un corpus de 5 M de segmentos)

Autismo es una condición neurológicos permanente que manifiestos durante la primera infancia, independientemente del género, la raza o la condición socioeconómica.

Marian (neuronal, entrenado con un corpus de 18 M de segmentos)

Es una condición neurofísica permanente que se manifiesta durante la primera infancia, independientemente del género, la raza o la situación socioeconómica.

La traducción automática neuronal

Ventajas

- Consigue mejor calidad que las técnicas predecesoras.
- Ofrece traducciones más fluidas.
- En general capta mejor el sentido de las oraciones.
- Consigue buena calidad incluso para lenguas ”alejadas”.
- La calidad es en general suficiente para flujo: **T.A. + postedición**

Inconvenientes

- Requiere corpus paralelos de gran tamaño.
- Para el entrenamiento requiere hardware específico: GPU.
- Trabaja con un vocabulario limitado (precisa subwords).
- Tiene tendencia a ofrecer traducciones más cortas que el original.
- Puede producir ”alucinaciones”.
- El resultado es muy fluido pero no siempre es preciso.

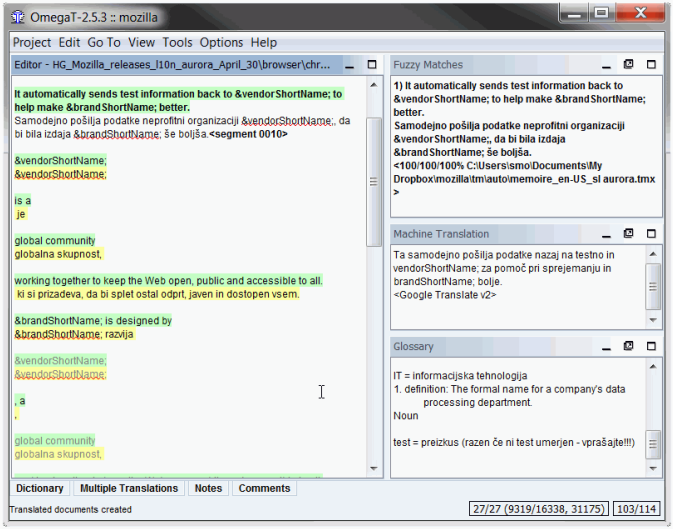
Acceso a la T.A.N.

- Servicios comerciales: Google Translate, Microsoft Bing Translator, DeepL...
- Crear nuestros propios sistemas

Acceso mediante API

- Nos interasa que nuestras aplicaciones se conecten automáticamente al sistema.
- Uso típico: en una herramienta de traducción asistida por ordenador. Consulta:
  - Memorias de traducción
  - Bases de datos terminológicacs
  - Sistemas de traducción automática

Ejemplo: OmegaT


OmegaT-2.5.3 :: mozilla
Project Edit Go To View Tools Options Help
Editor - HG_Mozilla_releases_l10n_aurora_April_30\browser\chr...
It automatically sends test information back to &vendorShortName; to help make &brandShortName; better.
Samodejno pošilja podatke neprofitni organizaciji &vendorShortName;, da bi bila izdaja &brandShortName; še boljša.<segment 0010>
&vendorShortName;
&vendorShortName;
is a
je
global community
globalna skupnost,
working together to keep the Web open, public and accessible to all.
ki si prizadeva, da bi splet ostal odprt, javen in dostopen vsem.
&brandShortName; is designed by
&brandShortName; razvija
&vendorShortName;
&vendorShortName;
, a
,
global community
globalna skupnost,
Fuzzy Matches
1) It automatically sends test information back to &vendorShortName; to help make &brandShortName; better.
Samodejno pošilja podatke neprofitni organizaciji &vendorShortName;, da bi bila izdaja &brandShortName; še boljša.
<100/100/100% C:\Users\smo\Documents\My Dropbox\mozilla\tm\auto\memoire_en-US_sl aurora.tmx >
Machine Translation
Ta samodejno pošilja podatke nazaj na testno in vendorShortName; za pomoč pri sprejemanju in brandShortName; bolje.
<Google Translate v2>
Glossary
IT = informacijska tehnologija
1. definition: The formal name for a company's data processing department.
Noun
test = preizkus (razen če ni test umerjen - vprašajte!!!)
Dictionary
Multiple Translations
Notes
Comments
Translated documents created
27/27 (9319/16338, 31175)
103/114

Sistemas comerciales: Ventajas

- Buena calidad.
- No requieren inversión en desarrollo.
- Acceso mediante API a precios muy reducidos:
  - Google Translate:
    - Primeros 500K char. Gratis
    - 20$ cada 1M char.
  - Microsoft Bing Translator:
    - Primeros 2M char. Gratis
    - 10$ cada 1M char.
  - DeepL:
    - TAO: 19.95 €/mes
    - API: 4.99 €/mes + 20€ cada 1M char.

Don Quijote: 2.659.336 char.

Sistemas comerciales: Inconvenientes

- ¡Confidencialidad!: enviamos datos a servidores externos
- Sistemas genéricos, no especializados
- Algunos (Google T., Microsoft Bing) permiten personalizarlos: difícil y caro.

Crear nuestros propios sistemas: Ventajas

- Motores personalizados: temáticas, clientes...
- No hay coste asociado por volumen.
- No hay problemas de confidencialidad: funcionan en nuestros propios ordenadores
- Se pueden conseguir mejores resultados para motores especializados.

Toolkits libres

- Moses (*www.statmt.org/moses/*)
- Marian (*https://github.com/marian-nmt/marian*)
- OpenNMT (*http://opennmt.net*)
- ModernMT (*www.modernmt.eu/*)
- Sockeye (*https://github.com/awslabs/sockeye*)
- JoeyNMT (*https://github.com/joeynmt/joeynmt*)

Alumni

Corpus paralelos

Corpus paralelos para el entrenamiento

- Opus Corpus (*http://opus.nlpl.eu/*)
- Memorias de traducción
- Alineación automática de documentos

Alumni

El proyecto MTUOC

Proyecto MTUOC: objetivos

Diversos *toolkits* SMT y NMT.

Problemas:

- Sistema Operativo: Linux
  - Máquinas virtuals
  - Windows Linux Subsystem
- Conocimientos técnicos.
- Hardware (memoria SMT; GPU NMT)
- Integración con herramientas TAO y flujos de trabajo profesionales

Objetivo:

- Facilitar el entrenamiento, uso e integración de sistemas TA neuronales (y estadísticos)

MTUOC: componentes

- Módulos de Python (tokenization, truecasing...)
- Scripts (preprocesamiento de corpus, archivos de configuración para el entrenamiento...)
- Scripts y programa de evaluación
- MTUOC-Server
- MTUOC-Translator
- Máquina virtual MTUOC
- Motores de traducción entrenados (principalmente Marian)

Preprocesamiento

TOKENIZACIÓN - DETOKENIZACIÓN:

They’re the teacher’s best students.

They ’re the teacher ’s best students .

They ■’re the teacher ■’s best students ■.

TRUECASING

La Primavera llegará pronto a Barcelona .

la primavera llegará pronto a Barcelona .

LIMPIEZA

Eliminar los segmentos demasiado largos.

TRATAMIENTO DE EXPRESIONES NUMÉRICAS

The income has reached 1,234 €

The income has reached @NUM@ €

The income has reached 1 ■, ■2 ■3 ■4 €

SUBWORDS

this subsystem will we reevaluated using molecular statitics .

this subsystem will we re@@ evaluated using molecular st@@ ati@@ tics .

—

Este sub@@ sistema será re@@ evaluado usando esta@@ díst@@ itica molecular .

Este subsistema será reevaluado usando estadístitica molecular .

SENTENCEPIECE

the Tuscany protects natural areas of exceptional value starting from Casentino national Park, extended to the Apennine ridges inhabited by deer and wolves, while its analogue in the archipelago comprises not only the emerged part of the Tyrrhenian Islands but also great biological diversity grounds.

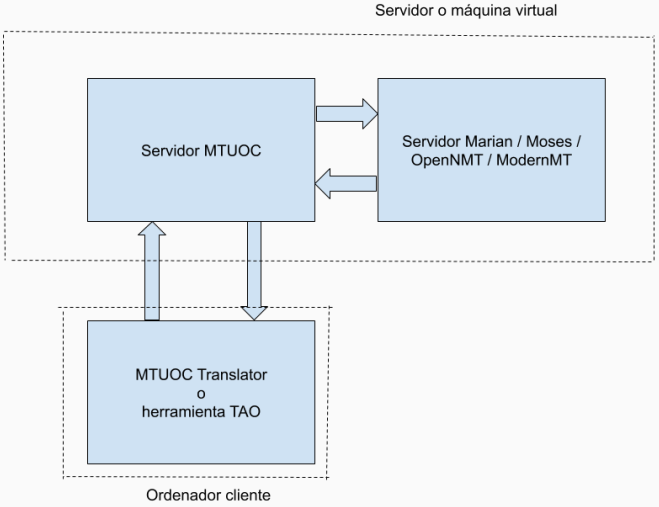
—

<s> _the _Tuscany _protects _natural _areas _of _exceptional _value _starting _ from _Cas entino _national _Park , _extended _to _the _Ap enn ine _rid ges _inh abited _by _deer _and _wolves , _while _its _analogue _in _the _archipelago _comprises _not _only _the _emerged _part _of _the _T yr r hen ian _Islands _bu t _also _great _biological _diversity _grounds . </s>

MTUOC-server

- Puede comportarse como MTUOC, Moses, OpenNMT, NMTWizard y ModernMT
- Recibe una petición para traducir del cliente
- Preprocesa el segmento
- Lo envía al servidor de traducción
- Recibe un segmento traducido y lo de-procesa
- Envía la traducción al cliente

MTUOC-server

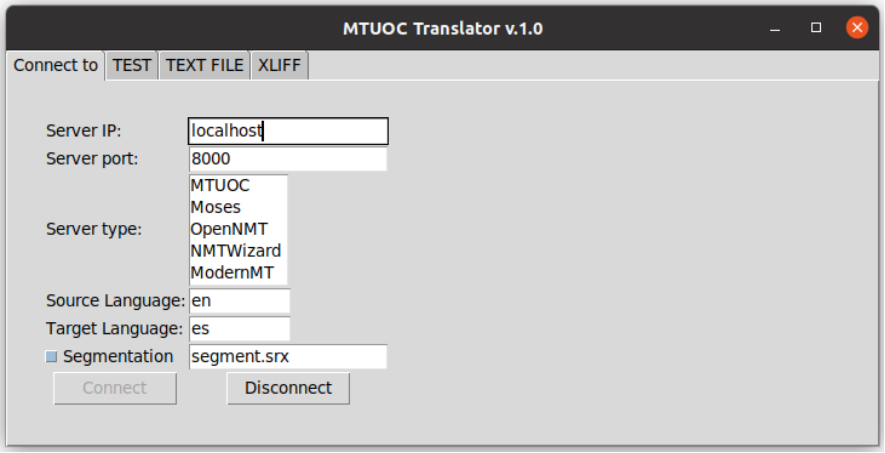


MTUOC-server

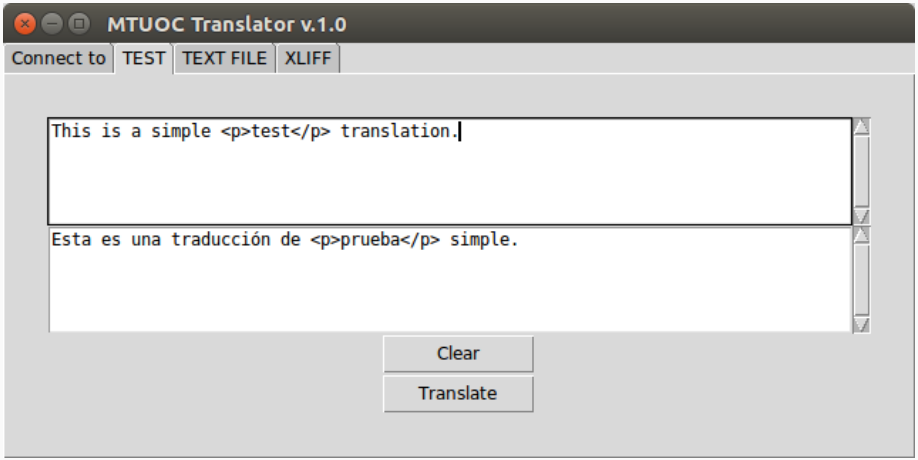
```
[2019-06-22 09:10:57] [config] type: s2s
[2019-06-22 09:10:57] [config] ulr: false
[2019-06-22 09:10:57] [config] ulr-dim-emb: 0
[2019-06-22 09:10:57] [config] ulr-dropout: 0
[2019-06-22 09:10:57] [config] ulr-keys-vectors: ""
[2019-06-22 09:10:57] [config] ulr-query-vectors: ""
[2019-06-22 09:10:57] [config] ulr-softmax-temperature: 1
[2019-06-22 09:10:57] [config] ulr-trainable-transformation: false
[2019-06-22 09:10:57] [config] version: v1.7.6 02f4af4 2018-12-12 18:51:10 -0800
[2019-06-22 09:10:57] [config] vocabs:
[2019-06-22 09:10:57] [config]   - vocab-en.yml
[2019-06-22 09:10:57] [config]   - vocab-es.yml
[2019-06-22 09:10:57] [config] word-penalty: 0
[2019-06-22 09:10:57] [config] workspace: 512
[2019-06-22 09:10:57] [config] Loaded model has been created with Marian v1.7.6
02f4af4 2018-12-12 18:51:10 -0800
[2019-06-22 09:10:57] [data] Loading vocabulary from JSON/Yaml file vocab-en.yml
[2019-06-22 09:10:57] [data] Loading vocabulary from JSON/Yaml file vocab-es.yml
[2019-06-22 09:10:57] [memory] Extending reserved space to 512 MB (device cpu0)
[2019-06-22 09:10:57] Loading scorer of type s2s as feature F0
[2019-06-22 09:10:57] Loading model from model.npz
[2019-06-22 09:11:07] Server is listening on port 8080
MTUOC server listening at port: 8000
```

UOC Alumni

MTUOC-Translator



MTUOC-Translator



Pasos de traducción

This is a simple test translation.

This is a simple <tag0>test</tag0> translation.

This is a simple test translation.

this is a simple test translation .

<s> _this _is _a _simple _test _translation _. </s>

Best translation 0 : <s> _esta _es _una _traducción_de _prueba _simple . ||| 0-0 1-1 2-2 3-3 4-5 4-7 5-6 6-4 7-8 8-9

<s> _esta _es _una _traducción _de _prueba _simple .

<s> _esta _es _una _traducción _de _<tag0> _prueba _</tag0> _simple .</s>

Esta es una traducción de prueba simple.

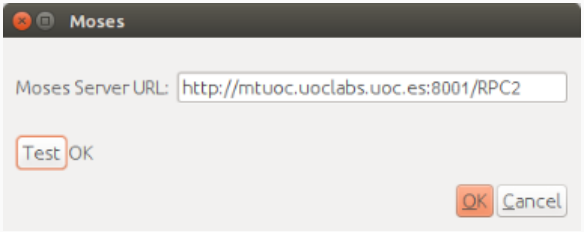
Conexión con herramientas TAO

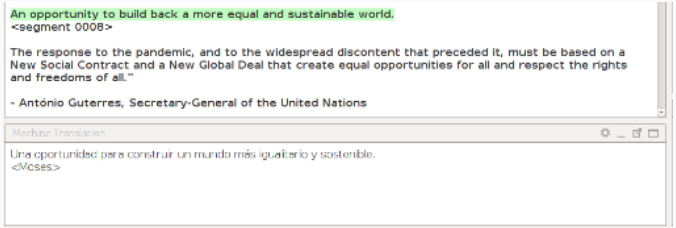
- SDL Trados
- OmegaT
- Okapi Tools (Raibow, tikal...)
- ...

tikal.sh -x -seg segment.srx file.docx -mmt http://192.168.1.39:8004

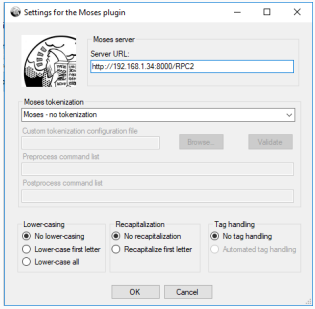
Alumni

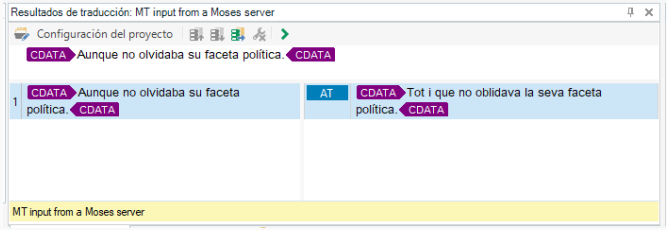
Conexión con OmegaT



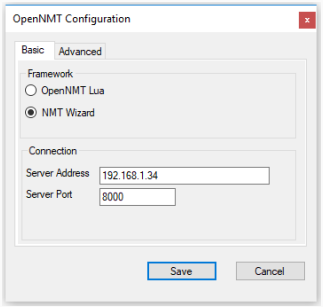

An opportunity to build back a more equal and sustainable world.
<segment 0008>
The response to the pandemic, and to the widespread discontent that preceded it, must be based on a New Social Contract and a New Global Deal that create equal opportunities for all and respect the rights and freedoms of all."
- António Guterres, Secretary-General of the United Nations
Machine Translation
Una oportunidad para construir un mundo más igualitario y sostenible.
<Moses>

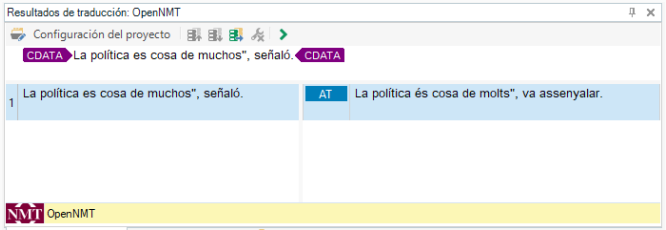
Conexión con SDL-Trados (2017) (como Moses)





Conexión con SDL-Trados (2017) (como NMTWizard)

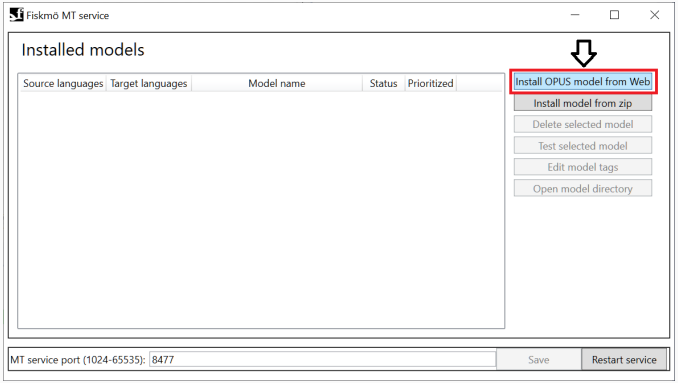




Proyectos similares

- Opus-MT: *https://github.com/Helsinki-NLP/Opus-MT*
- Fiskmo MT: *https://github.com/Helsinki-NLP/OPUS-CAT*
- Huggingface MarianMT: *https://huggingface.co/transformers/model_doc/marian.html*

FiskmoMT


Fiskmö MT service
Installed models
Source languages
Target languages
Model name
Status
Prioritized
Install OPUS model from Web
Install model from zip
Delete selected model
Test selected model
Edit model tags
Open model directory
MT service port (1024-65535):
8477
Save
Restart service

MTUOC: motores libres disponibles

python3 MTUOC-downloader.py -l

MTUOC-Moses-00-generic-eng-cat MTUOC-Moses-MultiUN-5M-eng-spa
MTUOC-Marian-00-generic-eng-cat MTUOC-Marian-00-generic-cat-eng
MTUOC-Marian-COVID19-eng-spa MTUOC-Marian-UNPC-eng-spa
MTUOC-Marian-UNPC-fra-spa MTUOC-Marian-UNPC-rus-spa

python3 MTUOC-downloader.py -e MTUOC-Marian-COVID19-eng-spa

https://huggingface.co/Helsinki-NLP

python3 MTUOC-download-HuggingFace.py –source en –target es

config-server.yaml

SL: ru TL: es server$_{t}$*ype* :
MTUOConeofMTUOC, *Moses*, *ModernMT*, *OpenNMT*, *NMTWizardport* : 8010

ONMT$_{u}$*rl*$_{r}$*oot* : */translator*

python3 MTUOC-server-HuggingFace.py

Información sobre MTUOC

- Github: *https://github.com/aoliverg/MTUOC*
- Artículo en The Conversation (divulgativo): *https://theconversation.com/como-crear-nuestro-propio-sistema-de-traduccion-automatica-ne*
- Paper EAMT: *https://www.aclweb.org/anthology/2020.eamt-1.55/*
- Vídeo divulgativo: *https://youtu.be/KLpHAkIAXrw*

UOC Alumni

Conclusiones

Conclusiones

- Los *toolkits* con licencias libres nos permiten crear nuestros propios sistemas T.A.N.
- Ciertas dificultades.
- Proyectos como MTUOC facilitan el entrenamiento, uso e integración de T.A.N.
- Principales ventajas:
  - Valor añadido a nuestros servicios.
  - Buena calidad de traducción.
  - No dependencia tecnológica.
  - Confidencialidad.

UOC Alumni

¡Muchas gracias por vuestra atención!

Antoni Oliver

aoliverg@uoc.edu

Alumni