

DESCRIPCIÓ DEL PROCÉS D'ALINEACIÓ AUTOMÀTICA DE DOCUMENTS

0. Objectius

El present document té com a objectiu descriure el procés d'alineació automàtica de documents i servir com a manual d'usuari de les eines que cal fer servir.

Entenem per alineació automàtica de documents el procés pel qual es relacionen els segments d'un determinat document o conjunt de documents amb els corresponents segments de la traducció del document o conjunt de documents a una altra llengua sense cap mena d'intervenció humana.

El formats que es pretenen tractar amb les eines que descriurem son el doc (Microsoft Word, RTF (Rich Text Format), html i txt.

Les eines que s'han desenvolupat estan programades en Perl i per tant són multiplataforma, però per poder fer la conversió dels formats doc, rtf i html necessita tenir instal·lat el Word en la versió Windows i per tant només funcionaran sota aquest sistema operatiu.

1. Descripció bàsica del procés

L'alineació automàtica de documents té dues fases diferenciades:

- Pretractament dels documents. Aquesta fase inclou els processos de:
 - Transformació del format original a txt (transformació que no cal si el format d'entrada és txt)
 - Segmentació dels documents
- Alineació automàtica pròpiament dita. L'algorisme que fem servir necessita que els fitxers d'entrada estiguin en format txt i prèviament segmentats

Les dues fases es fan amb el mateix programa *alinUOC*, però es poden cridar, si es considera necessari, de manera independent.

2.- Descripció dels programes necessaris

Per poder fer servir el alineador automàtic *alinUOC* cal tenir en un directori qualsevol els següents programes (que es distribueixen tant en font Perl com en executable per a Windows):

- Alinia-longitud
- Alinia-longitud-diccionari
- Alinuoc
- Construeix-model-one

- D2T
- Filtratge-final
- Filtratge-inicial
- Html2txt

Si es vol fer servir els fitxers Perl cal tenir instal·lat un intèrpret de Perl (es pot descarregar un de www.activestate.com). Si es fan servir els fitxer executables per Windows no caldrà tenir instal·lat l'intèrpret de Perl.

3. Requisits dels fitxers a alinear

Els documents a alinear han d'acomplir els següents requisits

- Format: els documents han d'estar en format:
 - Txt
 - Doc
 - Rtf
 - Html
- Ubicació i nom dels fitxers:
 - Tots els fitxers a alinear han de tenir un nom que acabi en `_` + codi de llengua (per exemple: `document1_cat.doc`). El codi de llengua pot ser qualsevol, però si es tracta de documents en català, castellà o anglès aconsellem fer servir `_cat`, `_spa` i `_eng` respectivament. D'aquesta manera les opcions del segmentador es seleccionaran automàticament).
 - Els documents originals i traduïts han de tenir el mateix nom amb codis de llengua diferents.
 - Només poden haver dos codis de llengua diferents

4.- Utilització del programa *alinuoc*

El programa *alinUOC* funciona en línia de comandes. Per veure l'ajuda del programa cal fer:

```
alinuoc -h
```

Si fem això se'ns mostra el següent missatge:

```
Ajuda;
Programa ALINUOC: alineació de documents). Aquest programa està basat
en el
    bilingual sentence aligner de Moore. Els paràmetres són:
--do (string): Directori original. Obligatori.
--dd (string): Directori destí. Obligatori.
--np: Només pretractament. Opcional.
--na: Només alineació. Opcional.,n--ef: Elimina del directori els
fitxers que no tenen correspondència. Opcional.
--nd: Neteja directori. Esborra tots els arxius auxiliars del
directori. Opcional.
--th (real): Thereshold. Opcional. Si no s'especifica cap es pren 0.5.
```

--fs (string): Fitxer de sortida on es guarda la memòria de traducció. Opcional.
--fna (string): Fitxer que conté els segments no alineats. Es creen dos, corresponents a cada una de les llengües que estan en joc. Opcional.
--help: mostra aquesta ajuda (també es mostra fent --h)
Final de l'ajuda

Exemple d'utilització:

Volem alinear els documents que es troben al directori C:/documentsalineacio, deixant els resultats al directori C:/resultatalineacio. La memòria de traducció volem que es digui meo.txt i els fitxers de segments no alineats que es digui noalineats

```
alineacio --do=C:/documentsalineacio --dd=resultatalineacio --nd --fs=meo.txt --fna=noalineats
```

Amb això s'executarà la alineació i es crearà la memòria. Per saber si el resultat de l'alineació ha tingut èxit es pot comparar la mida de la memòria resultant amb la mida dels fitxers de segments no alineats. Si l'alineació no ha funcionat bé, es pot repetir especificant un valor de threshold diferent amb l'opció --th. Per evitar tornar a fer el preprocessament dels documents es pot repetir el procés d'alineació fent servir l'opció -na.

5.- Distribució dels programes i instal·lació

Per instal·lar els programes només cal descomprimir l'arxiu zip a un directori qualsevol.

ANNEX

Reproducció de la llicència de l'algorisme d'alineació automàtica de Moore:

BILINGUAL SENTENCE ALIGNER

This Microsoft Research end user license agreement ("MSR-EULA") is a legal agreement between you and Microsoft Corporation ("Microsoft" or "we") for the software or data identified above, which may include source code, and any associated materials, text or speech files, associated media and "online" or electronic documentation (together, the "Software").

By installing, copying, or otherwise using the Software, found at <http://research.microsoft.com/downloads>, you agree to be bound by the terms of this MSR-EULA. If you do not agree, do not install, copy or use the Software. The Software is protected by copyright and other intellectual property laws and is licensed, not sold.

If Software is in object code:

* Upon your agreement to the terms below, you may install one copy of the Software on your personal computer and use such copy at no charge for your non-commercial research or non-commercial teaching purposes, only in an academic or other noncommercial research setting. In return, we ask that you agree to the following:

* **NO TRANSFER RIGHTS:** That you will not copy, sell, rent, lease, distribute, sublicense, assign, or otherwise transfer (including by loan or gift) the Software and will not attempt to modify it, or to reverse engineer or decompile it, except and only to the extent authorized.

* To leave in place all copyright notices and licensing information that you might find in the Software.

* That you will not use the Software in a live operating environment where it may be relied upon to perform in the same manner as a commercially released product, or with data that has not been sufficiently backed up.

* That any feedback about the Software provided by you to us is voluntarily given, and Microsoft shall be free to use the feedback as it sees fit without obligation or restriction of any kind, even if the feedback is designated by you as confidential.

* **NO WARRANTIES WHATSOEVER:** That the Software comes "AS IS", with all faults and with no warranties, conditions or representations. The implied warranties of merchantability and fitness for a particular purpose, and any warranty against interference with your enjoyment of the Software or against infringement, do not apply to the Software. The entire risk as to satisfactory quality, performance, accuracy, and effort concerning the Software is assumed by you. There is no warranty that this Software will fulfill any of your particular purposes or needs.

* That we have no duty of reasonable care or lack of negligence, and we are not obligated to (and will not) provide technical support for the Software.

* That we will not be liable for any damages, including those known as direct, indirect, special, consequential, or incidental damages related to the Software, this MSR-EULA, or under any legal theory (such as negligence), to the maximum extent that overriding applicable law permits.

* That if you sue or threaten to sue anyone over patents that you think may apply to the Software, or if you breach this MSR-EULA in any way, your license to the Software ends automatically.

* That this MSR-EULA shall be construed and controlled by the laws of the State of Washington, USA, without regard to conflicts of law.

If Software is in data form or source code, these additional rights and restrictions apply:

* You may do anything you want with the Software source code or data for non-commercial research or non-commercial teaching purposes free of charge, provided that you agree to the following:

* To make freely available to others the source code or data of any modifications or additions you make to the Software, and any related documentation, solely and exclusively under the same terms as this License.

* That Microsoft is granted back, without limitations, the rights to reproduce, install, use, modify, distribute and transfer your modifications to the Software source code or data.

Copyright (c) Microsoft Corporation. All rights reserved.