On Cross-lingual Plagiarism Analysis using a Statistical Model

Alberto Barrón-Cedeño, Paolo Rosso,
David Pinto and Alfons Juan

Universidad Politécnica de Valencia

July, 2008

Overview

- Introduction
- Probabilistic decision rule
- Plagiarism probabilistic model
- Maximum likelihood estimation
- Experiments
- Conclusions
- Further Work

Introduction:CLiPA

Plagiarise

To robe credit of another person's work; in text it means including text fragments from an author without giving him the corresponding credit

Introduction:CLiPA

Plagiarise

To robe credit of another person's work; in text it means including text fragments from an author without giving him the corresponding credit

Cross-lingual perspective

A text fragment in *one language* is considered a plagiarism of a text in *another language* if their contents are considered semantically similar no matter they are written in different languages and the corresponding citation or credit is not included

Introduction:CLiPA

Plagiarise

To robe credit of another person's work; in text it means including text fragments from an author without giving him the corresponding credit

Cross-lingual perspective

A text fragment in *one language* is considered a plagiarism of a text in *another language* if their contents are considered semantically similar no matter they are written in different languages and the corresponding citation or credit is not included

We call this task **CLiPA** (Cross-Lingual Plagiarism Analysis)

Introduction: Related work

Pouliquen, et. al. (2003)

Search for document translations using the Eurovoc thesaurus (http://europa.eu/eurovoc/)

ES	DE	EN
abandono escolar	vorzeitiger Schulabgang	dropout
abastecimiento	Versorgung	supply
abastecimiento energético	Energieversorgung	energy supply
abogado	Rechtsanwalt	barrister
abono	Düngemittel	fertiliser
abono del suelo	Bodendüngung	soil conditioning
abono orgánico	organischer Dünger	organic fertiliser
abono químico	chemischer Dünger	chemical fertiliser
aborto	Abtreibung	abortion
aborto ilegal	illegale Abtreibung	illegal abortion
aborto terapéutico	indizierte Abtreibung	therapeutic abortion
Abruzos	Abruzzen	Abruzzi
absentismo	Fernbleiben von der Arbeit	absenteeism
abstencionismo	Wahlenthaltung	abstentionism
Abu Dabi	Abu Dhabi	Abu Dhabi
abuso de confianza	Veruntreuung	breach of trust
abuso de derecho	Rechtsmissbrauch	misuse of a right
abuso de información privilegiada	Insidergeschäft	insider trading
abuso de poder	Amtsmissbrauch	abuse of power
acceso a la educación	Zugang zur Bildung	access to education
acceso a la información	Informationszugang	access to information
acceso a la información comunitaria	Zugang zu Gemeinschaftsinformationen	access to Community information
acceso a la justicia	Zugang zur Rechtspflege	access to the courts
acceso a la profesión	Zugang zum Beruf	access to a profession
acceso al empleo	Zugang zur Beschäftigung	job access

Introduction: Preliminary approach

Potthast, et. al. (2008)

First “real” approach to CLiPA

We still not have too much information

(but we know that we are not alone)

Probabilistic decision rule

$x_1, x_2, \cdots x_V$	Suspicious text fragments	(L_a)
$y_1, y_2, \cdots, y_W$	Original text fragments	(L_b)

Probabilistic decision rule

$x_1, x_2, \cdots x_V$ Suspicious text fragments (L_a)

$y_1, y_2, \cdots, y_W$ Original text fragments (L_b)

$$y_i^*(x) = \underset{y=y_i \cdots y_w}{\operatorname{argmax}} p(y \mid x)$$

where $p(y \mid x)$ is an appropriate plagiarism (translation) probabilistic model.

Plagiarism probabilistic model

$p(y \mid x)$ is modelled by using the IBM M1 statistical alignment model (Brown, et. al., 1990)

$$p(y \mid x) = \prod_{j=1}^{|x|} \sum_{i=0}^{|y|} p(i \mid j, |y|)\ p(x_j \mid y_i)$$

with: $p(i \mid y, |y|) = \frac{1}{|y|+1}$

$p(x_j \mid y_i)$ (statistical dictionary)

Maximum likelihood estimation

Using the EM algorithm:

$$p(w|v)^{(k+1)} = \frac{N(w,v)}{\sum_{w'} N(w',v)}$$

$$N(w|v) = \sum_n \frac{p(w \mid v)^{(k)}}{\sum_{j'} p(w \mid x_{nj'})^{(k)}} \sum_{i=1}^{|y_n|} \sum_{j=0}^{|x_n|} \delta(y_{ni}, w)\, \delta(x_{nj}, v)$$

for all $v \in \mathcal{X}$ and $w \in \mathcal{Y}$; where $\delta(a,b)$ is the Kronecker delta function

Maximum likelihood estimation

Example:

L_a	L_b	
$a_1\, a_2\, a_3 \cdots$	$b_4\, b_6 \cdots$	a_2 is equally related to $b_4, b_6 \cdots$
$a_4\, a_5\, a_2 \cdots$	$b_1\, b_3\, b_6\, b_2 \cdots$	relation of a_2 and b_6 is stronger

Experiments

Corpus

- 5 original fragments (*En*)
  - *Intrinsic plagiarism analysis deals with the detection of plagiarised sections within a document d, without comparing d to extraneous sources*

Experiments

Corpus

- 5 original fragments (*En*)
  - Y *Intrinsic plagiarism analysis deals with the detection of plagiarised sections within a document d, without comparing d to extraneous sources*
- 9 human plagiarised fragments (*Sp*)
  - X_h El analisis del plagio intrínseco tiene que ver con la detección de secciones plagiadas de un documento d, sin comparar d con fuentes externas

Experiments

Corpus

- 5 original fragments (*En*)
  - Y *Intrinsic plagiarism analysis deals with the detection of plagiarised sections within a document d, without comparing d to extraneous sources*
- 9 human plagiarised fragments (*Sp*)
  - X_h El analisis del plagio intrínseco tiene que ver con la detección de secciones plagiadas de un documento d, sin comparar d con fuentes externas
- 5 translated fragments (*Sp*)
  - X_t El analisis de plagio intrinseco trata con la detección de secciones plagiadas dentro de una d de documento, sin comparar la d a fuentes extrañas

Experiments

Corpus

- 5 original fragments (*En*)
  - Y *Intrinsic plagiarism analysis deals with the detection of plagiarised sections within a document d, without comparing d to extraneous sources*
- 9 human plagiarised fragments (*Sp*)
  - X_h El analisis del plagio intrínseco tiene que ver con la detección de secciones plagiadas de un documento d, sin comparar d con fuentes externas
- 5 translated fragments (*Sp*)
  - X_t El analisis de plagio intrinseco trata con la detección de secciones plagiadas dentro de una d de documento, sin comparar la d a fuentes extrañas
- 46 original fragments originally written in Spanish

Experiments

Corpus

Are these cases different enough to the model calculation?

Experiments

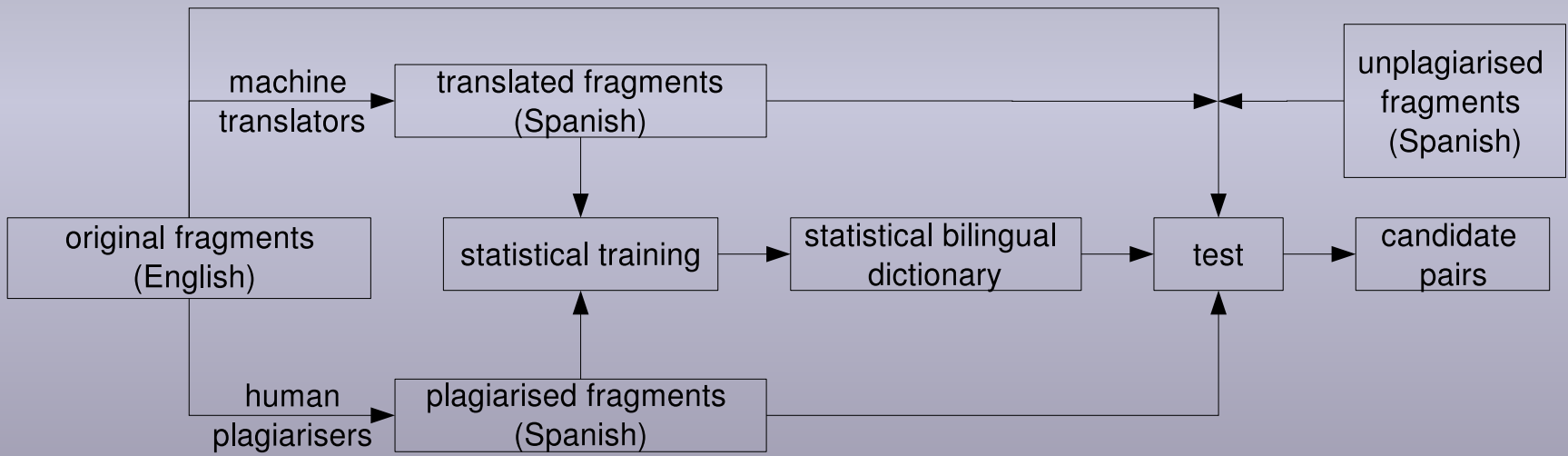
Corpus

Are these cases different enough to the model calculation?

$J_\delta(Y_1, *)$	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9
t_1	.50	.58	.58	.50	.52	.74	.58	.44	.37
t_2	.54	.59	.48	.45	.52	.73	.52	.41	.41
t_3	.51	.60	.60	.54	.51	.75	.61	.45	.43
t_4	.56	.64	.60	.56	.54	.76	.63	.48	.43
t_5	.67	.74	.73	.70	.66	.78	.67	.65	.63

Experiments

Experiment description



Experiments

Example

X_1 *El análisis del plagio intrinseco tiene que ver con la detección de secciones plagiadas de un documento d, sin comparar d con fuentes externas*

Z_1 *Hipótesis La perplejidad de un fragmento perteneciente a un escritor con respecto a otro, será mayor que la de dos documentos escritos por el mismo autor. Aquellos párrafos que tengan mayor perplejidad sera los mejores candidatos a ser fragmentos plagiados.*

Experiments

Example

X_1 *El análisis del plagio intrinseco tiene que ver con la detección de secciones plagiadas de un documento d, sin comparar d con fuentes externas*

Z_1 *Hipótesis La perplejidad de un fragmento perteneciente a un escritor con respecto a otro, será mayor que la de dos documentos escritos por el mismo autor. Aquellos párrafos que tengan mayor perplejidad sera los mejores candidatos a ser fragmentos plagiados.*

fragment	result	decision
X_1	$p(Y_5\|X_1) = 33.1 \cdot 10^{-5}$	
	$p(Y_i\|X_1) \longrightarrow 0 \forall i \neq 5$	plagiarised from Y_5

Experiments

Example

X_1 *El análisis del plagio intrinseco tiene que ver con la detección de secciones plagiadas de un documento d, sin comparar d con fuentes externas*

Z_1 *Hipótesis La perplejidad de un fragmento perteneciente a un escritor con respecto a otro, será mayor que la de dos documentos escritos por el mismo autor. Aquellos párrafos que tengan mayor perplejidad sera los mejores candidatos a ser fragmentos plagiados.*

fragment	result	decision
X_1	$p(Y_5\|X_1) = 33.1 \cdot 10^{-5}$	
	$p(Y_i\|X_1) \longrightarrow 0 \forall i \neq 5$	plagiarised from Y_5
Z_1	$p(Y_i\|Z_1) \approx 0 \forall i$	original

Conclusions

1. We presented a method for cross-lingual plagiarism analysis based on a statistical bilingual dictionary

Conclusions

1. We presented a method for cross-lingual plagiarism analysis based on a statistical bilingual dictionary
2. The order of the words is not relevant and we are able to find good candidates even when the plagiarised fragment has been modified

Conclusions

1. We presented a method for cross-lingual plagiarism analysis based on a statistical bilingual dictionary
2. The order of the words is not relevant and we are able to find good candidates even when the plagiarised fragment has been modified
3. We believe that this technique is valuable resource for the CLiPA task.

Further work

- To validate our results on a bigger corpus

(Unfortunately, the construction of a cross-lingual corpus with the required characteristics seems to be by itself a sufficiently difficult task)

Thank you

Alberto Barrón-Cedeño, Paolo Rosso,
David Pinto and Alfons Juan

{lbarron, prosso, dpinto, ajuan}@dsic.upv.es