Information Retrieval Research at INF-UFRGS

Viviane Moreira
Instituto de Informática – UFRGS – Brazil
viviane@inf.ufrgs.br

Outline

- Motivation
- Previous Work
  - Cross-Language Information Retrieval
  - Cross-Language Plagiarism Detection
  - Mapping Wikipedia Infoboxes Across Languages
- Ongoing Work
  - Crawling the Web for Parallel Corpora
  - Finding Missing Cross-Language Links at Wikipedia

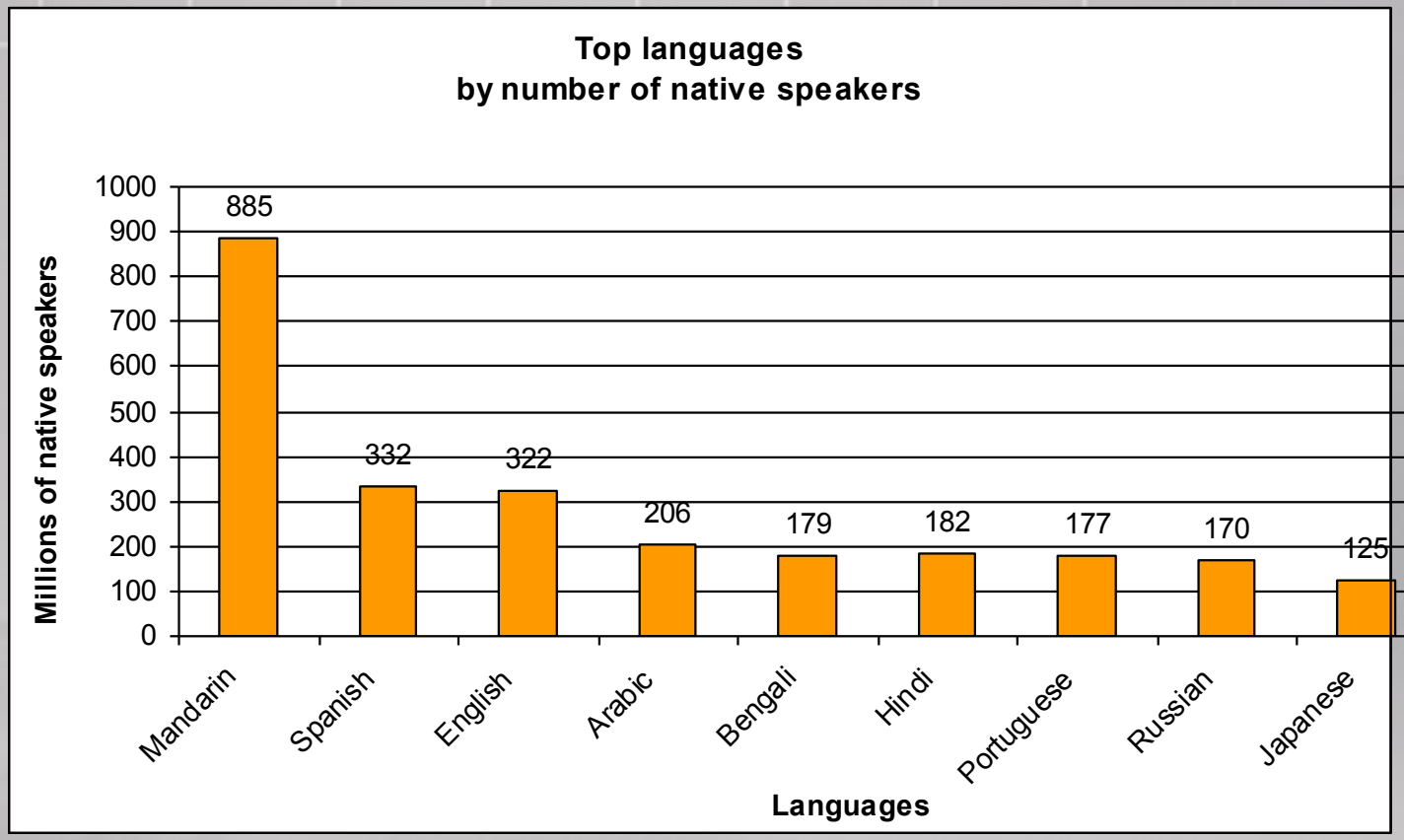
UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Cross-Language Information Retrieval

Motivation

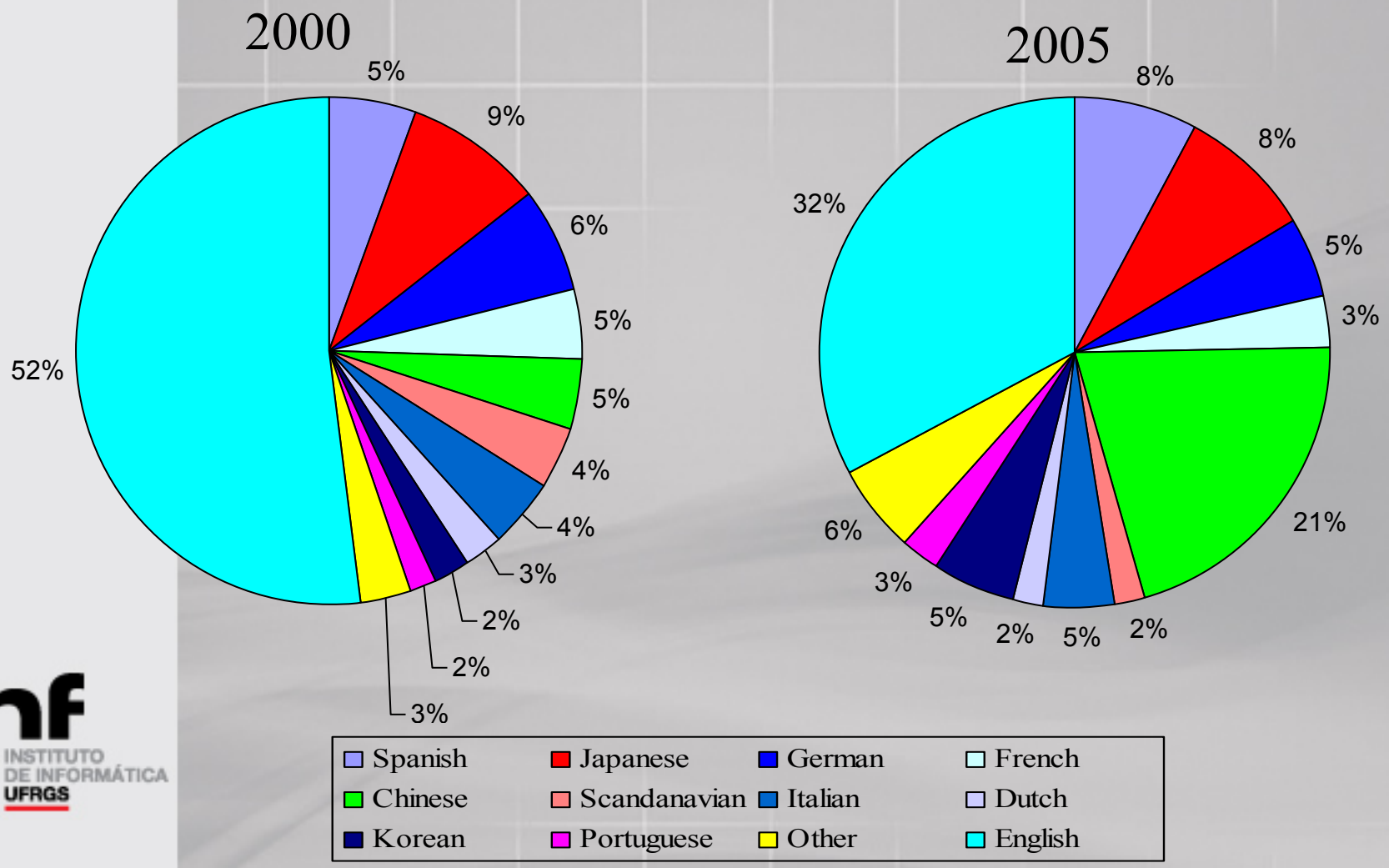
.inf INSTITUTO DE INFORMÁTICA UFRGS

Languages by number of native speakers



Source:http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

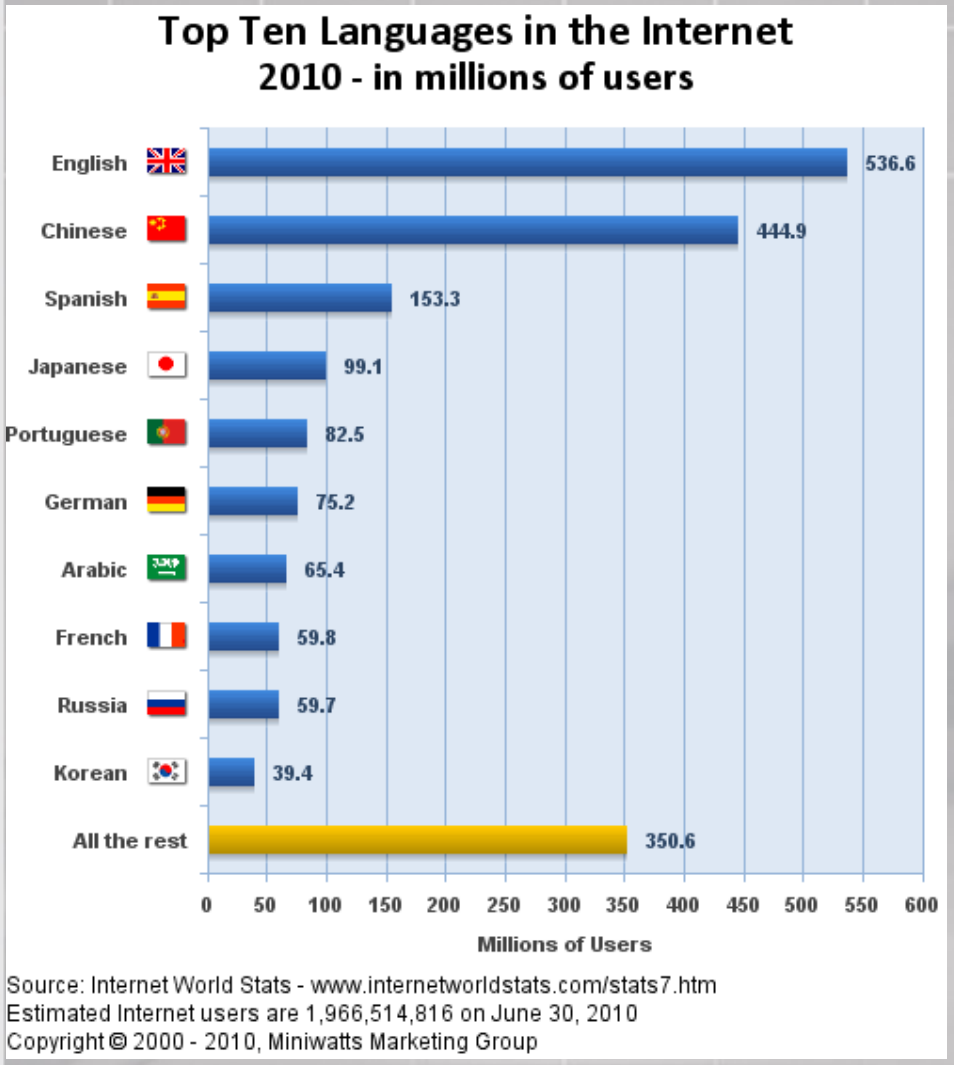
Internet Users



Source: Miguel Ruiz http://informatics.buffalo.edu/faculty/ruiz/teaching/Seminars/Cross-Language_information_Retrieval.ppt

2010

Internet Users



Source: http://www.internetworldstats.com/stats7.htm

2010

Internet Users

Top Ten Languages in the Internet

Top Ten Languages Used in the Web (Number of Internet Users by Language)					
TOP TEN LANGUAGES IN THE INTERNET	Internet Users by Language	Internet Penetration by Language	Growth in Internet (2000 - 2011)	Internet Users % of Total	World Population for this Language (2011 Estimate)
English	565,004,126	43.4 %	301.4 %	26.8 %	1,302,275,670
Chinese	509,965,013	37.2 %	1,478.7 %	24.2 %	1,372,226,042
Spanish	164,968,742	39.0 %	807.4 %	7.8 %	423,085,806
Japanese	99,182,000	78.4 %	110.7 %	4.7 %	126,475,664
Portuguese	82,586,600	32.5 %	990.1 %	3.9 %	253,947,594
German	75,422,674	79.5 %	174.1 %	3.6 %	94,842,656
Arabic	65,365,400	18.8 %	2,501.2 %	3.3 %	347,002,991
French	59,779,525	17.2 %	398.2 %	3.0 %	347,932,305
Russian	59,700,000	42.8 %	1,825.8 %	3.0 %	139,390,205
Korean	39,440,000	55.2 %	107.1 %	2.0 %	71,393,343
TOP 10 LANGUAGES	1,615,957,333	36.4 %	421.2 %	82.2 %	4,442,056,069
Rest of the Languages	350,557,483	14.6 %	588.5 %	17.8 %	2,403,553,891
WORLD TOTAL	2,099,926,965	30.3 %	481.7 %	100.0 %	6,930,055,154

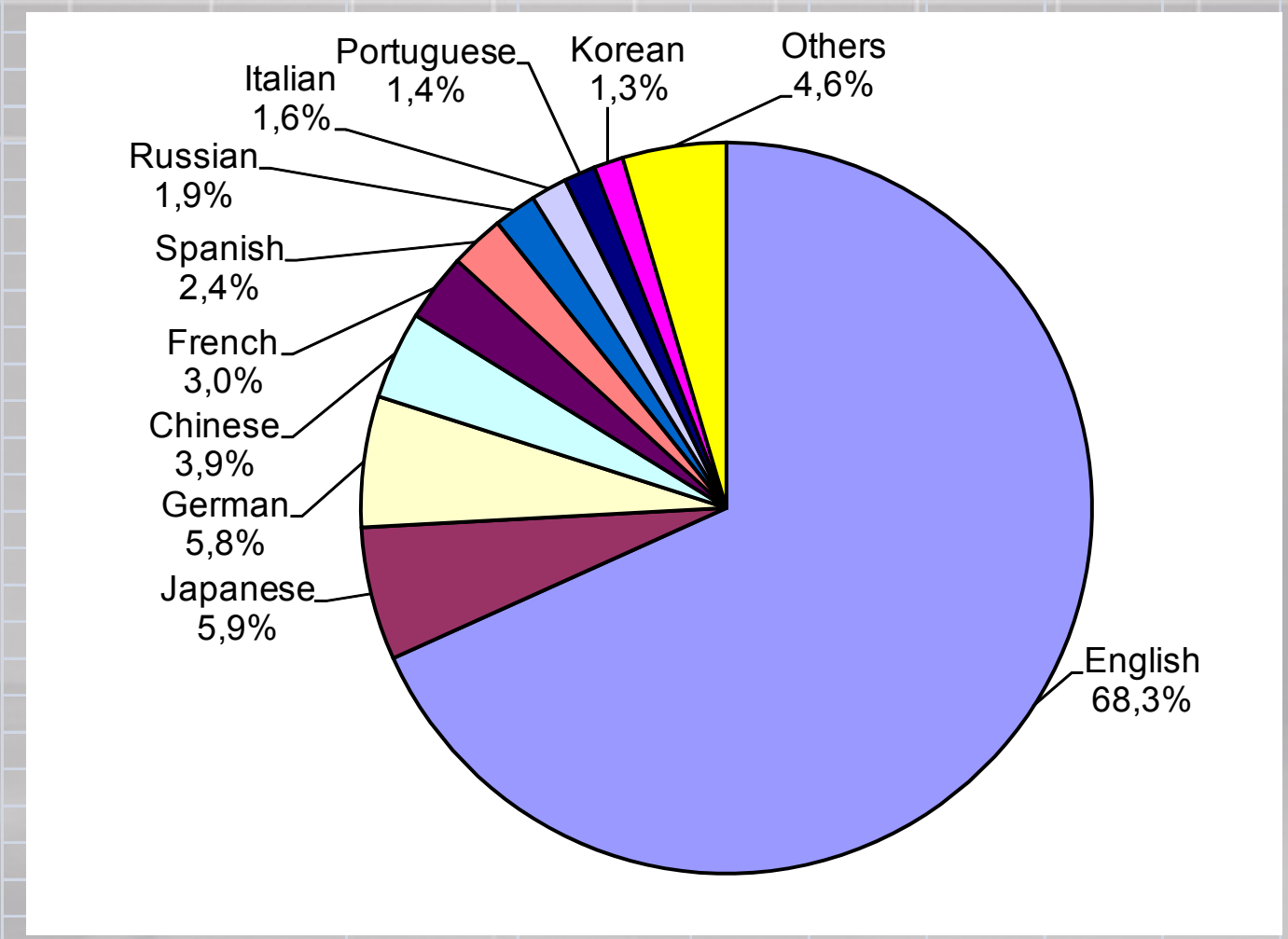
NOTES: (1) Top Ten Languages Internet Stats were updated for May 31, 2011. (2) Internet Penetration is the ratio between the

Millions of Users

Source: Internet World Stats - www.internetworldstats.com/stats7.htm
Estimated Internet users are 1,966,514,816 on June 30, 2010

Source: http://www.internetworldstats.com/stats7.htm

Web's content by language



Fonte: Global Reach Internet Statistics http://global-reach.biz/globstats/refs.php3

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

CLIR using Association Rules

MSc Dissertation by André Geraldo

UFRGS – 2009

.inf INSTITUTO DE INFORMÁTICA UFRGS

Context & Motivation

- Cross-Language Information Retrieval (CLIR): is the retrieval of documents in one natural language, based on a query formulated in another natural language
- Corpus-based strategies
- Challenge: efficiently analyse large text collections and generate good mappings between terms across languages

Association Rules

- An association rule is an implication $X \Rightarrow Y$ where $X = \{x_1, x_2, \ldots, x_n\}$, and $Y = \{y_1, y_2, \ldots, y_m\}$ are sets of items
- Example:
  "90% of customers who purchase bread also purchase milk"
- Support and Confidence

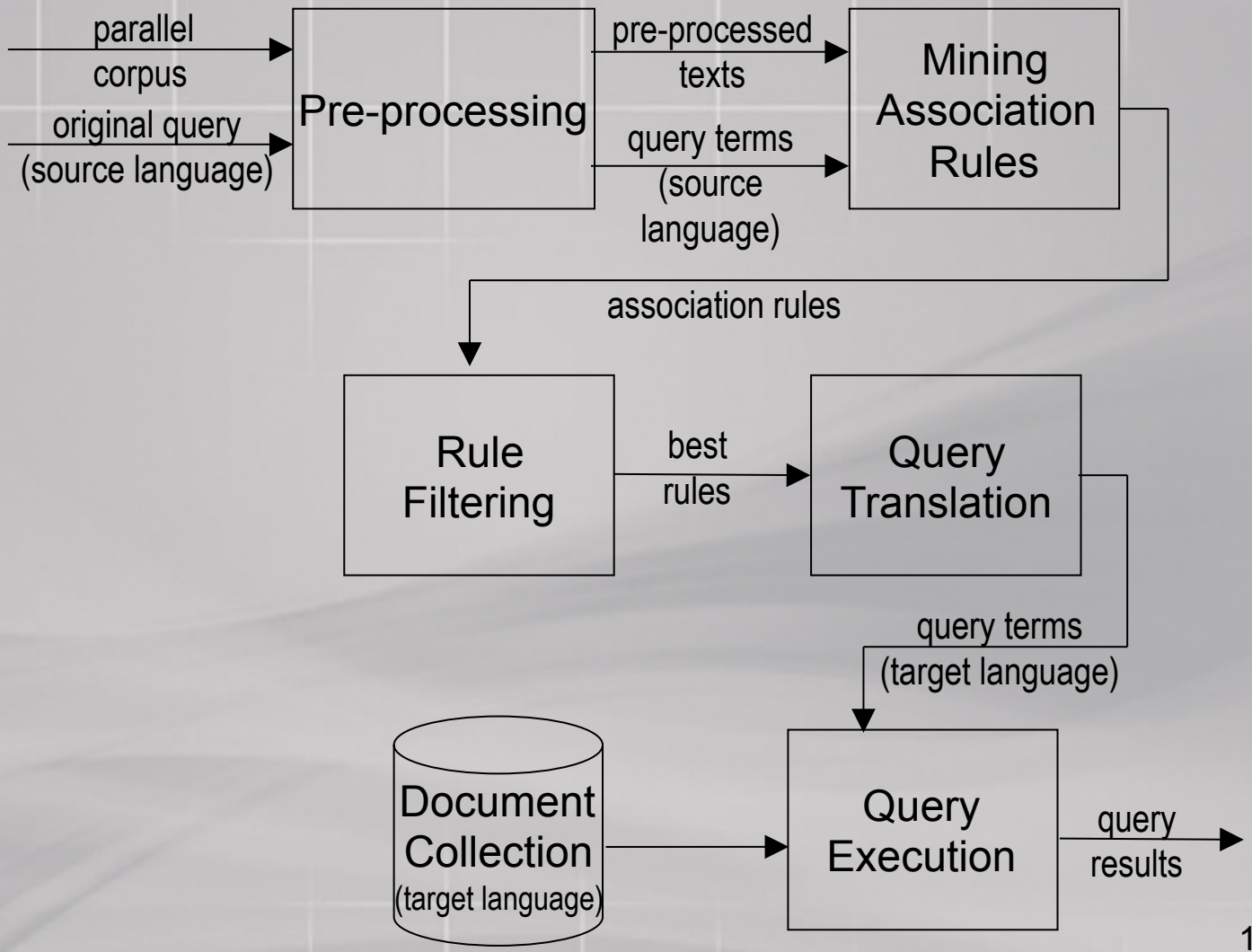
$$\sup(X \Rightarrow Y) = \frac{n(X \bigcup Y)}{N} \qquad conf(X \Rightarrow Y) = \frac{n(X \bigcup Y)}{n(X)}$$

- Task: generate all rules that have support and confidence greater than predefined thresholds

Our proposal

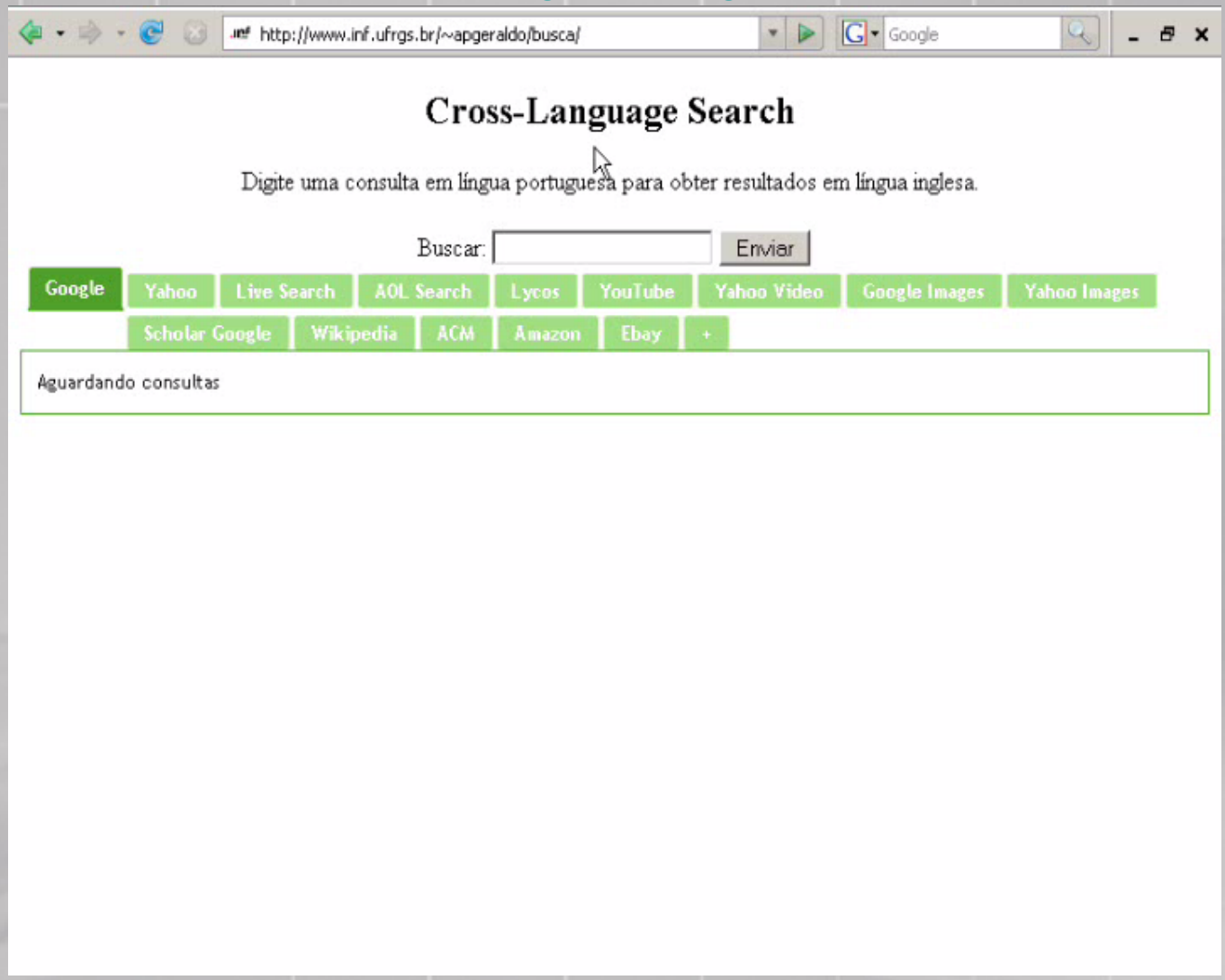
Map the problem of finding association rules between items in a market-basket scenario to the problem of finding cross-linguistic equivalents between a pair of languages over a parallel corpus

Proposed Approach



Demo

http://www.inf.ufrgs.br/~apgeraldo/busca/



Summary

- Proposed the use of algorithms for mining association rules to the problem of finding term translations for CLIR.
- Method requires a parallel corpus to serve as the basis for the mining process.
- Carried out experiments in which queries in Finnish and in Portuguese are used to retrieve documents in English.
- Our results are comparable to the monolingual baseline and to the results of Machine Translation Systems applied to the purpose of translating queries.
- The combination of Machine Translation and our Association Rule-based approach even surpassed the monolingual baseline for the Portuguese topics.

Publications

- Geraldo, A.P., V.P. Moreira, and M.A. Gonçalves, ***On-Demand Associative Cross-Language Information Retrieval*, in *String Processing and Information Retrieval***, *16th International Symposium (SPIRE)*, 2009, Springer: Saariselkä, Finland. p. 165-173.

- Geraldo, A.P. and V.M. Orengo, *UFRGS@CLEF2008: **Using Association rules for Cross-Language Information Retrieval***, in *Evaluating Systems for Multilingual and Multimodal Information Access (CLEF)* 2008, Springer: Aarhus, Denmark. p. 66-74.

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Cross-Language Plagiarism Detection

MSc Dissertation by Rafael Pereira

UFRGS – 2010

.inf INSTITUTO DE INFORMÁTICA UFRGS

Introduction

- Plagiarism is the use of another person's written work without acknowledging the source
- Several types of plagiarism
  - copying another's work word-for-word
  - paraphrasing the text in order to disguise the offense
  - cross-language content translation
- A study by McCabe with over 80,000 students in the US and Canada found that 36% of undergrad students and 24% of postgrad students admit to have copied or paraphrased sentences from the Internet without referencing them.

McCabe, D.L., *Cheating among college and university students: A North American perspective.* International Journal for Educational Integrity, 2005.

Motivation

- Maurer et al. surveyed plagiarism detection systems and found that none of the available tools support search for cross-language plagiarism.

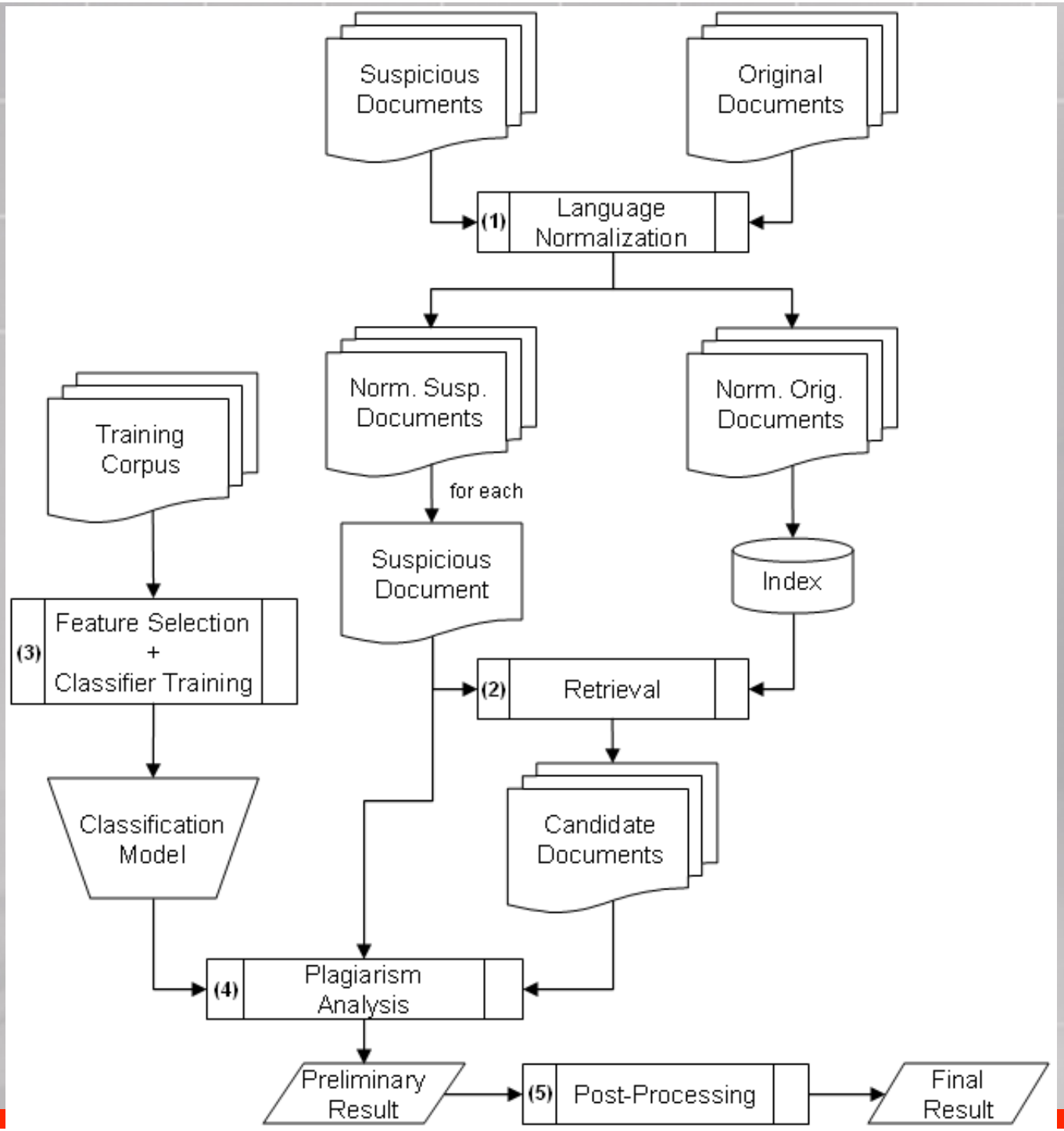
- The increasing availability of textual content in many languages, and the evolution of automatic translation can potentially make this type of plagiarism more common.

Maurer, H., F. Kappe, and B. Zaka, *Plagiarism - A Survey.* Journal of Universal Computer Science, 2006: p. 1050-1084.

The Task

- Detect the plagiarized passages in the suspicious documents and their corresponding text fragments in the source documents even if the documents are in different languages.

- Known as **Extrinsic** plagiarism analysis


Suspicious Documents
Original Documents
(1) Language Normalization
Norm. Susp. Documents
Norm. Orig. Documents
Training Corpus
for each
Suspicious Document
Index
(3) Feature Selection + Classifier Training
(2) Retrieval
Classification Model
Candidate Documents
(4) Plagiarism Analysis
Preliminary Result
(5) Post-Processing
Final Result

Summary

- Took part on the PAN campaign for the evaluation of plagiarism detection systems
- Built two bilingual corpora for testing plagiarism detection systems (Portuguese-English and French-English) http://www.inf.ufrgs.br/~viviane/eclapa.html
- Cross-language experiment achieved 86% of the performance of the monolingual baseline

Publications

- Pereira, R., V.P. Moreira and R. Galante. ***A New Approach for Cross-Language Plagiarism Analysis*** in *CLEF*. Padua, Italy 2010.

- Pereira, R., V.P. Moreira and R. Galante. UFRGS@PAN2010: Detecting External Plagiarism. In: *Proceedings of the PAN 2010 Lab on Uncovering Plagiarism, Authorship, and Social Software Misuse*, 2010.

Schema Matching for Wikipedia Infoboxes

Work done during my sabbatical at the University of Utah (09/2010 to 08/2011)

In collaboration with Juliana Freire and Thanh Nguyen

Motivation

- With ~18 million articles and ~10 million pageviews a month, Wikipedia has consolidated itself as an important source of information
- ... in special multilingual information
- It also has structured information (infoboxes)

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf
INSTITUTO DE INFORMÁTICA
UFRGS

The Last Emperor

Directed by	Bernardo Bertolucci
Produced by	Jeremy Thomas
Written by	Mark Peploe Bernardo Bertolucci
Starring	John Lone Joan Chen Peter O'Toole
Music by	Ryuichi Sakamoto
Cinematography	Vittorio Storaro
Editing by	Gabriella Cristiani
Studio	Hemdale Film Corporation
Distributed by	Columbia Pictures
Release date(s)	October 23, 1987 (Italy) November 18, 1987
Running time	160 minutes
Country	China United Kingdom Italy
Language	English Mandarin Chinese
Budget	$23.8 million[1]
Box office	$43,984,230[2]

Wikipedia Infoboxes for the film The Last Emperor in EN and PT

O Último Imperador

O Último Imperador (PT/BR)

Reino Unido / Itália
França / China
1987 • cor • 165 min

Produção

Direção	Bernardo Bertolucci
Roteiro	Mark Peploe / Bernardo Bertolucci
Elenco original	John Lone Joan Chen Peter O'Toole Ryuichi Sakamoto
Género	drama biográfico / épico
Idioma original	inglês / mandarim / japonês

IMDb: (inglês) (português)

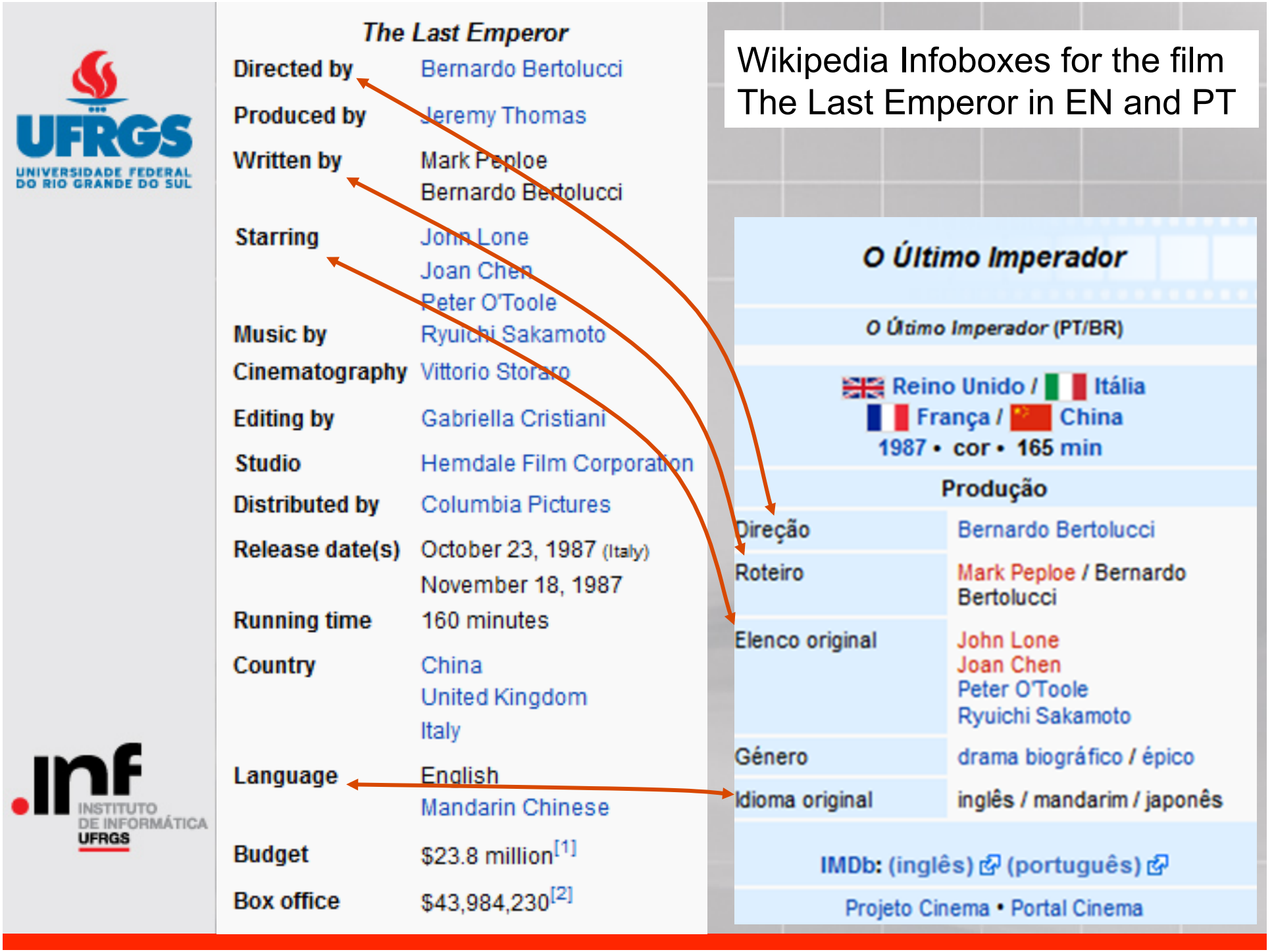
Projeto Cinema • Portal Cinema

Wikipedia Infoboxes for the film The Last Emperor in EN and PT

The Last Emperor	
Directed by	Bernardo Bertolucci
Produced by	Jeremy Thomas
Written by	Mark Peploe Bernardo Bertolucci
Starring	John Lone Joan Chen Peter O'Toole
Music by	Ryuichi Sakamoto
Cinematography	Vittorio Storaro
Editing by	Gabriella Cristiani
	Italy
Language	English Mandarin Chinese
Budget	$23.8 million[1]
Box office	$43,984,230[2]

O Último Imperador	
O Último Imperador (PT/BR)	
Reino Unido / Itália França / China 1987 • cor • 165 min	
Produção	
Direção	Bernardo Bertolucci
Roteiro	Mark Peploe / Bernardo Bertolucci
Elenco original	John Lone Joan Chen Peter O'Toole Ryuichi Sakamoto
Género	drama biográfico / épico
Idioma original	inglês / mandarim / japonês
IMDb: (inglês) (português)	
Projeto Cinema • Portal Cinema	

Problem: Find mappings between attributes in infoboxes in different languages


Wikipedia Infoboxes for the film
The Last Emperor in EN and PT
The Last Emperor
Directed by
Bernardo Bertolucci
Produced by
Jeremy Thomas
Written by
Mark Peploe
Bernardo Bertolucci
Starring
John Lone
Joan Chen
Peter O'Toole
Music by
Ryuichi Sakamoto
Cinematography
Vittorio Storaro
Editing by
Gabriella Cristiani
Studio
Hemdale Film Corporation
Distributed by
Columbia Pictures
Release date(s)
October 23, 1987 (Italy)
November 18, 1987
Running time
160 minutes
Country
China
United Kingdom
Italy
Language
English
Mandarin Chinese
Budget
$23.8 million[1]
Box office
$43,984,230[2]
O Último Imperador
O Último Imperador (PT/BR)
Reino Unido / Itália
França / China
1987 • cor • 165 min
Produção
Direção
Bernardo Bertolucci
Roteiro
Mark Peploe / Bernardo Bertolucci
Elenco original
John Lone
Joan Chen
Peter O'Toole
Ryuichi Sakamoto
Género
drama biográfico / épico
Idioma original
inglês / mandarim / japonês
IMDb: (inglês) (português)
Projeto Cinema • Portal Cinema

WikiMatch

- Different evidences of similarity
  - Attribute values
  - Link structure
  - Co-occurrence statistics
- Matched infoboxes in English, Portuguese, and Vietnamese
- High precision and recall, outperforming state-of-the-art techniques
- Case study demonstrating that the multilingual mappings derived lead to improvements in answer quality and coverage for structured queries

Publication

- Thanh Nguyen, Viviane Moreira, Huong Nguyen, Hoa Nguyen, Juliana Freire, **Multilingual Schema Matching for Wikipedia Infoboxes**. Proceedings of the VLDB Endowment. vol. 5 No. 2, p. 133-144, 2011

Outline

- Previous Work
  - ✓ Cross-Language Information Retrieval
  - ✓ Cross-Language Plagiarism Detection
  - ✓ Mapping Wikipedia Infoboxes Across Languages
- Ongoing Work
  - Crawling the Web for Parallel Corpora
  - Finding Missing Cross-Language Links at Wikipedia

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Crawling the Web for Multilingual Corpora

MSc Dissertation by Marcela Vieira

to be concluded in 2012

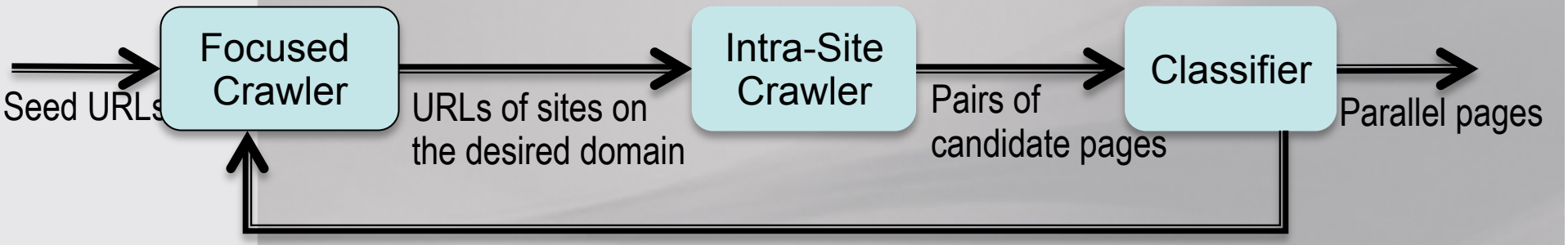
.inf INSTITUTO DE INFORMÁTICA UFRGS

Motivation

- Parallel Corpora are important resources for training CLIR and Machine Translation Systems
- The domain of the corpus plays an important role
- This is a scarce resource

The Task

- Given a language-pair and a domain (e.g. English-French and chemistry) crawl the web and find pairs of parallel pages to assemble a parallel corpus.



UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

Finding Missing Cross-Language Links at Wikipedia

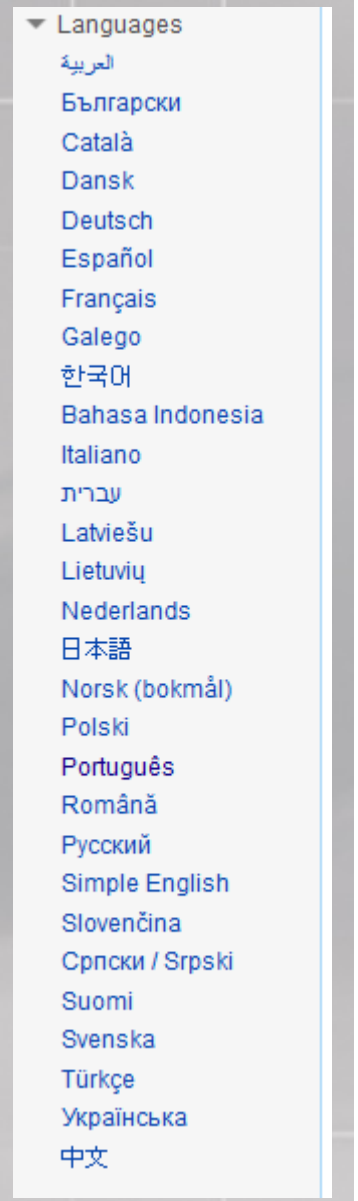
MSc Dissertation by Carlos E. M. Moreira

to be concluded in 2013

.inf INSTITUTO DE INFORMÁTICA UFRGS

Motivation

- Wikipedia has become an important source of comparable corpora – versions of the same article in different languages
- Cross-language links connect these different versions
- These links are manually created by Wikipedia editors
- Some of these links are missing



The Task

- Step 1) For a page without a link, find candidate correspondences in other languages
- Step 2) Use different evidences of similarity to determine whether a link should be placed

Other (not so related) research topics

- Monolingual Plagiarism Detection
- Stemming Algorithms
- Hidden Web Crawling

Thank You

- Questions
- Comments
- Ideas for cooperation

Please, email **viviane@inf.ufrgs.br**