A Parallel Corpus of Theses and Dissertations Abstracts

Felipe Soares, Gabrielli Harumi Yamashita, Michel Anzanello

felipe.soares@inf.ufrgs.br

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf
INSTITUTO DE INFORMÁTICA
UFRGS

Introduction

- Cross-language corpora is the basis of MT systems.
- Acquisition is not a trivial task.
- Already available parallel corpora:
  - Europarl (Koehn, 2005);
  - United Nations (UN) (Ziemski et al., 2016);
  - Patents (Utiyama and Isahara, 2007);
  - Movie subtitles (Zhang et al., 2014);
  - Books (Skadins et al. 2014);

Introduction

- Parallel corpora from scientific articles can be a valuable resource.
- Benefit for several NLP tasks:
  - Cross-language plagiarism detection;
  - Article indexing and classification;
  - Named entity recognition;
- Development of parallel corpora based on scientific texts already addressed by different authors:
  - Biomedical articles from PUBMED in 6 languages (Wu et al., 2011);
  - Annotated parallel corpus for biomedical NER (Kors et al, 2015);
  - Parallel corpus of biomedical abstracts from Scielo (Neves et al., 2016).

Introduction

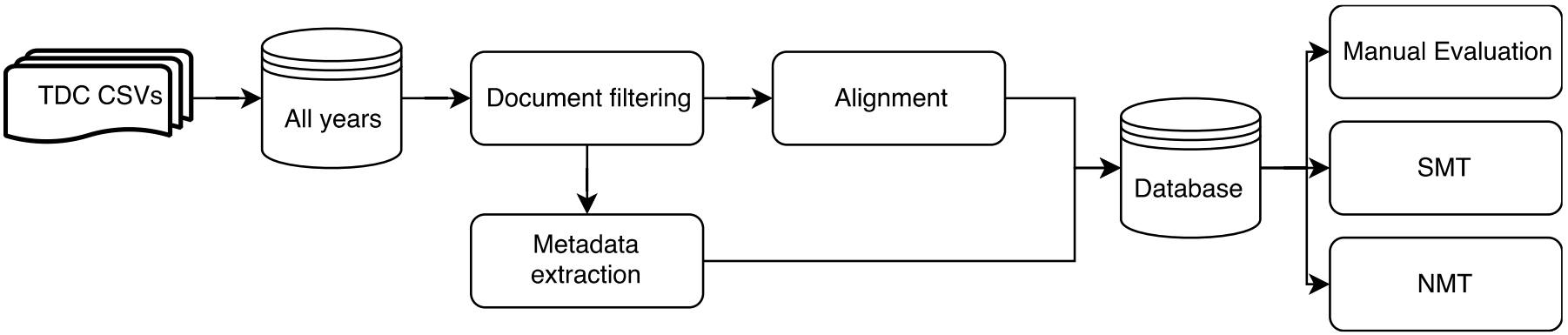
- In Brazil, CAPES is the governmental body responsible for overseeing post-graduate programs across the country.
- CAPES keeps track of every dissertation and thesis in a centralized database (TDC) under the open data initiative.
- **Present work:**
  - **Development of a parallel corpus of theses and dissertations abstracts in Portuguese and English.**
  - **Sentence aligned using Hunalign.**
  - **Spans the years from 2013 to 2016.**

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf UFRGS

Introduction

Material and Methods



- CSVs retrieved from CAPES website.
- Documents loaded in SQL database.
  - Documents without abstract in both PT and EN were removed.
  - Additional language checking.
- Alignment using Hunalign
- Experiments with SMT, NMT, and manual evaluation.

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf UFRGS

Issues with original data

Grid | Graph | Map | 1000 records | « | 1 – 100 | » | Search data ... | Go » | Filters

KEY...	DS_ABSTRACT	DS_RESUMO	IN_
OPA...	THE COPAIFERA PAUPERA (HERZOG) DWYER BELON...	A COPAIFERA PAUPERA (HERZOG) DWYER PERTENCE À FAMÍLI...	0
ETU...	OBJECTIVE: THIS STUDY AIMS TO EVALUATE THE M...	OBJETIVO: AVALIAR O ÍNDICE DE PERFORMANCE MIOCÁRDICA...	1
OOD...	THE AIM OF THE STUDY WAS TO INVESTIGATE THE ...	O OBJETIVO DO ESTUDO FOI VERIFICAR OS EFEITOS PROVOC...	1
OMI...	INTRODUCTION: THE ABDOMINOPLASTY IS THE THI...	INTRODUÇÃO: A ABDOMINOPLASTIA É O TERCEIRO PROCEDI...	1
RTI...	THIS STUDY APPROACHES ABOUT THEME RELIGIO...	ESTA DISSERTAÇÃO ABORDA AS DIVERGÊNCIAS ENTRE RELIGI...	1
NTH...	ABSTRACT A CENTRAL QUESTION IN ECOLOGY AN...	RESUMO UMA QUESTÃO CENTRAL EM ECOLOGIA E BIOLOGIA ...	0
ION ...	THE DEFOLIATING PESTS, ESPECIALLY CATERPILLA...	AS PRAGAS DESFOLHADORAS, PRINCIPALMENTE O COMPLEX...	0

- All text is presented in upper case letters, which jeopardizes more sophisticated approaches for sentence segmentation.

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf UFRGS

Sentence alignment

- Preprocessing:
  - Conversion to lower case
  - Language check to make sure that there was no misplacing of English abstracts in the Portuguese field.
  - Deletion of newline/carriage return (i.e. \n and \r)
- LF aligner as sentence alignment tool:
  - Wrapper based on Hunalign (Varga et al, 2007);
  - Pre-built dictionaries.
- Post-processing:
  - Removal of all non-aligned sentences;
  - Removal of sentences with fewer than three characters;

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf UFRGS

Evaluation

- Statistical Machine Translation (SMT):
  - Moses (Koehn et al., 2007);
  - Translation quality evaluated according BLEU score.
- Neural Machine Translation (NMT):
  - OpenNMT (Klein et al., 2017);
  - Translation quality evaluated according BLEU score.
- Translation was compared to Google Translate
- Manual evaluation of sentence alignment:
  - Randomly selection of 400 pairs of sentences;
  - If fully aligned -> “correct”;
  - If incompletely aligned -> “partial”;
  - If incorrectly aligned -> “no alignment”.

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf UFRGS

Results and Discussion

- The dataset is available in TMX format and SQLite database.
  - https://figshare.com/s/091fcaf8ad66a3304e90
- Metadata included:
  - Year
  - University
  - Title in Portuguese,
  - Type of document (dissertation or thesis)
  - Keywords in both languages
  - Knowledge areas and subareas according to CAPES
  - URL for the full-text PDF

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf UFRGS

Results and Discussion

- Corpus statistics

Knowledge Area	Docs	Sents	Tokens EN	Tokens PT
Health Sciences	38,221	224,773	5.46M	5.51M
Humanities	38,493	189,648	5.63M	5.54M
Applied Social Sciences	32,176	160,131	4.66M	4.60M
Agricultural Sciences	26,740	154,710	3.92M	3.92M
Engineering	27,074	149,888	3.87M	3.92M
Multidisciplinary	26,502	140,849	3.84M	3.81M
Exact and Earth Sciences	19,630	106,098	2.64M	2.66M
Biological Sciences	16,465	98,994	2.33M	2.34M
Linguistic and Arts	13,717	64,281	1.99M	1.96M
Total	239,018	1,289,372	34.35M	34.28M

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf UFRGS

Results and Discussion

- Datasets:
  - Sentences randomly split in training, development and test.
  - Approximately 13,000 sentences allocated to development and test.
- SMT models build based on Moses baseline system steps[1]
- NMT models build using OpenNMT Torch implementation:
  - 2-layer LSTM model
  - 500 hidden units in both encoder and decoder
  - 12 epochs
  - UNK words replaced by word in the input language

1- http://www.statmt.org/moses/?n=moses.baseline

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf UFRGS

Results and Discussion

- SMT and NMT Experiments.
- Comparison to Google Translate on the test set.
- NMT presented the best results, specially for EN->PT.

Partition	System	PT → EN	EN → PT
Dev	Moses	44.07	41.21
	OpenNMT	44.02	43.36
Test	Moses	43.85	41.05
	OpenNMT	**43.89**	**43.22**
	Google Translate	42.57	38.92

UFRGS
UNIVERSIDADE FEDERAL
DO RIO GRANDE DO SUL

.inf UFRGS

Results and Discussion

- Alignment Manual Evaluation
  - Correctly aligned; 82.30%
  - Partially aligned: 13.33%
  - No alignment: 4.35%
- Most of the problems in partial alignment were due to segmentations issues previous to the alignment, which wrongly split the sentences.

Portuguese	English
os dados foram comparados entre os grupos por anova de medidas repetida	data were compared by repeated measures anova. results: we found a significa
o estudo utilizará um software comercial para simular a peça	the study will use commercial software to simulate the piece with a number of different crack sizes and the
buscamos subsídios teóricos em autores que veem na reflexão e na pesquisa um grande potencial para o desenvolvimento d	we seek theoretical support in authors who see in reflection and research a great potential for

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf UFRGS

Conclusion

- Development of a parallel corpus of theses and dissertations abstracts in Portuguese and English.
- Corpus based on the CAPES TDC dataset from 2013 to 2016.
- Corpus tested on SMT, NMT experiments, and by manual evaluation of alignment.
- Experiments with best results than Google Translate, specially for the NMT models.
- Segmentation issues due to case folding.
- Future work:
  - Text classification and clustering
  - In-domain training for specific knowledge areas.

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf UFRGS

A Parallel Corpus of Theses and Dissertations Abstracts

Felipe Soares, Gabrielli Harumi Yamashita, Michel Anzanello

felipe.soares@inf.ufrgs.br

UFRGS
UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL

.inf
INSTITUTO DE INFORMÁTICA
UFRGS