NSOAMT

Lisb@20[20]

PORTUGAL 2020

UNIÃO EUROPEIA
Fundos Europeus Estruturais e de Investimento

PDM

New Search Only Approach to Machine Translation (NSOAMT)

Introduction

NSOAMT

Translation Automation: Bridging Language Barriers

- Developed over several years
- Deep learning-based tools dominate in the last year
- Driven by AI technology advancement and semantic abstraction

Introducing NSOAMT: A Research Project

- Addressing issues (e.g., inaccuracies) in existing approaches
- Focus on creating correspondence between native and target languages
- Utilizes incremental word indexing with shared semantic meaning

Key Assumption: Limited Vocabulary in Document Types

- Language style and word diversity are relatively restricted
- Enables instantaneous and accurate translations through indexing

Objectives

NSOAMT

Objectives

- Overcome the slowness and inaccuracy in translation semantics.
- Develop a solution based on incremental word indexing with shared semantic meaning.
- Enable a correspondence process between native language and translation.
- Achieve relatively instantaneous and accurate translation semantics.

Impact on Society

- Contribution to society's daily life and collective well-being.
- Simplify processes (administrative, legal, etc.) and reduce associated costs.
- Streamline relationships between citizens and companies.
- Benefit various sectors (information services, governments, multinationals, NGOs, etc.).

Objectives

NSOAMT

Expected Benefits

- Automation of translations for repeated phrases.
- Suggestions for translations of similar phrases.
- Immediate search in the translation memory.
- Creation of integrated terminology files with translations.
- Management of translation memory with external resources.
- Facilitation of placing translated text in final documents.
- Enhanced IT security of translated text.

Problem statement

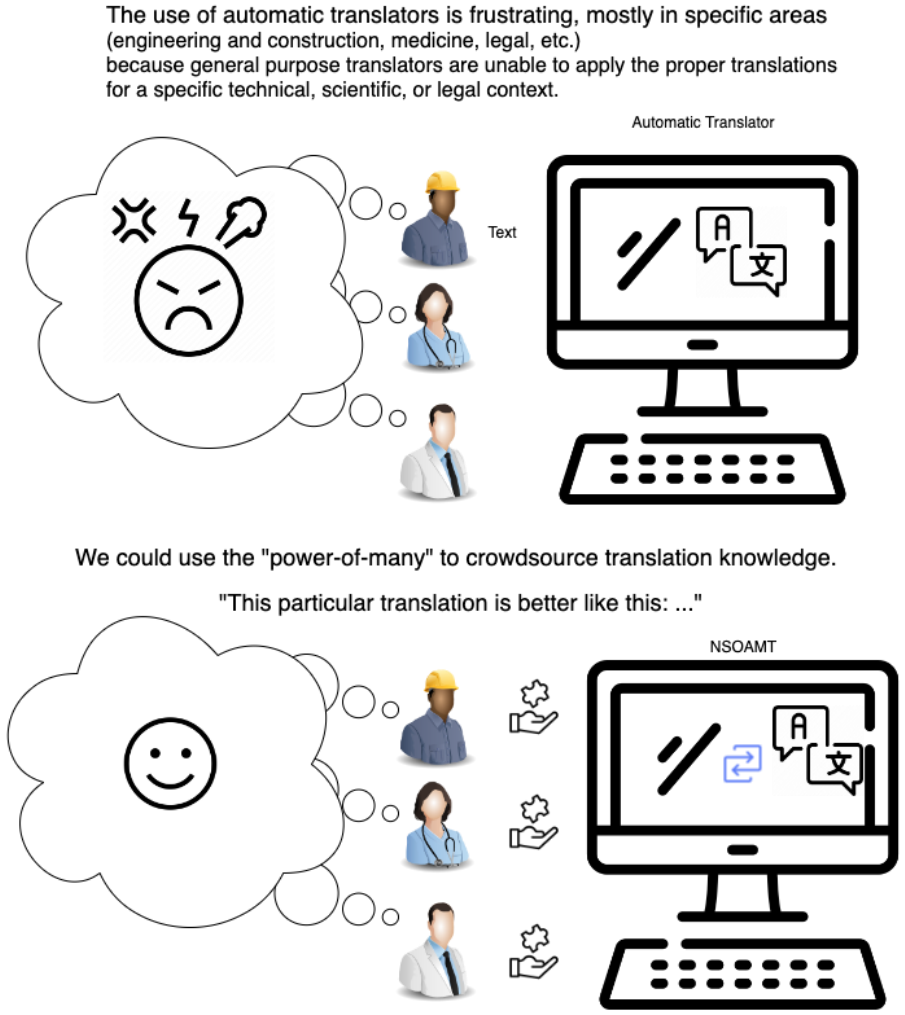
NSOAMT

Existing Translation Services

- General-purpose translation services available
- High-quality translations in specific domains often require human experts

Unlocking the Potential of Language

- Natural language sentences are word sequences
- Could domain expertise be crowd-sourced by quantifying and storing common sentences?
- Question: Can we bridge the gap in machine translation through crowd-sourcing?



Our Innovative Approach: NSOAMT

NSOAMT

Core Idea

- Create a correspondence between native language content and translation through an incremental word indexing approach.

Search-only Machine Translation

- The aim of this project is to develop a new and totally innovative approach called Search-only Machine Translation, based on the idea that "All meaningful sentences have already been written and translated", which will overcome the problems of current machine translation software.

How

- Development of algorithms based on cryptography techniques to efficiently convert a sentence into a single computational identifier.
- Development of task parallelization and data distribution algorithms to enable constant-time searches and joins.

Methodology

NSOAMT

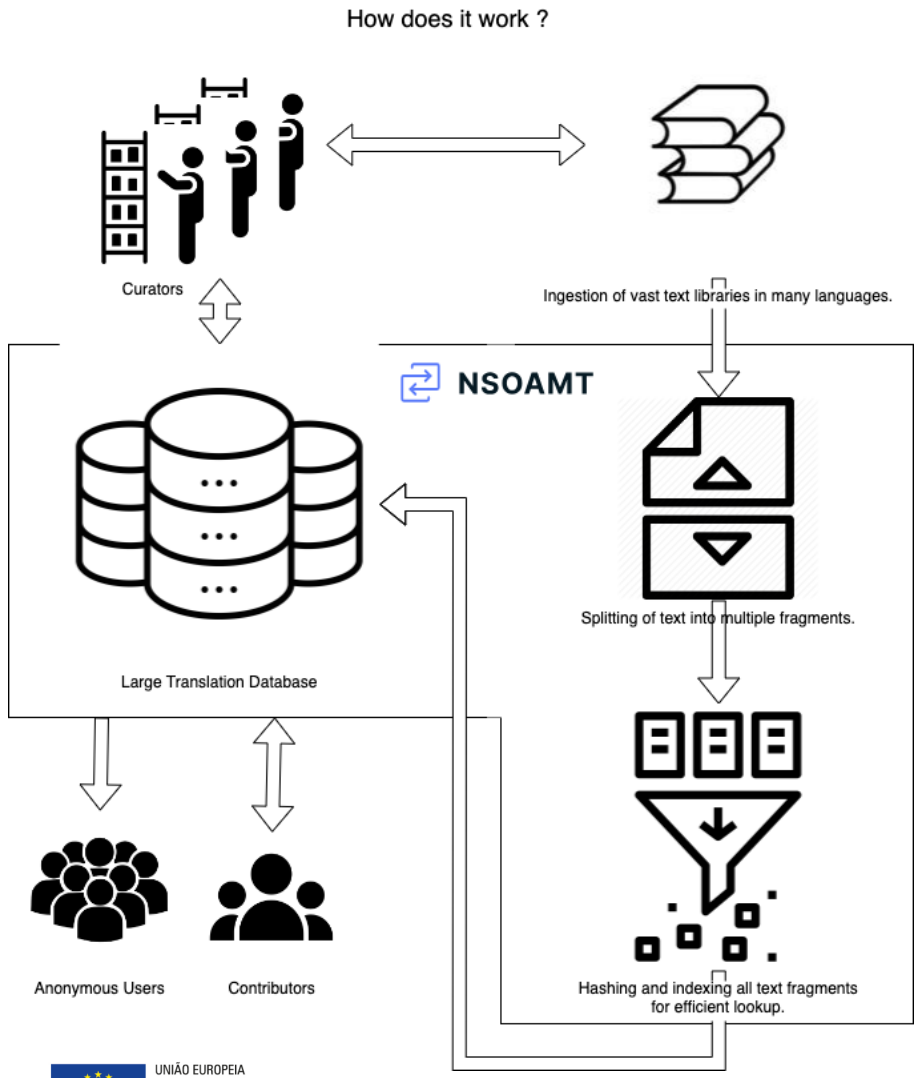
1. Import vast quantities of text documents, broken into sentences.

2. Identify common text fragments.

3. Crowdsource translation of common text fragments.

This methodology is only viable today due to:

- Connectivity of the Internet and World Wide Web linking organizations, institutions, and private initiatives, providing abundant text sources.
- Widespread availability of open-source libraries and tools like NLTK, facilitating rapid prototyping of NLP techniques.
- Advancements driven by Moore's Law, enabling cost-effective hardware capable of handling Terabytes of text data.



Implementation

NSOAMT

Text Sources

Having access to a diverse and ample collection of text sources is a critical success factor for project.

- https://eur-lex.europa.eu/ - Legislation documents from the European Union, available in 24 languages; HTML format.
- https://dumps.wikimedia.org/ - Wikipedia backup dumps. XML+Wikitext format.
- https://arxiv.org/ - Open-access scholarly articles. PDF format.1 Download was performed my mirroring tools, with articles organized in monthly folders. (Only the latest version of each article was ingested.)
- tBooks - Several plain text literature content, obtained from sources like https://www.gutenberg.org/, https://chroniclingamerica.loc.gov/, https://muse.jhu.edu/, https://market.cantook.com/, https://www.bookrix.com/, https://archive.org/, https://manybooks.net/, https://www.smashwords.com/, http://digital.library.upenn.edu/books/. Plain text (UTF-8) format. We call this source aggregate tBooks

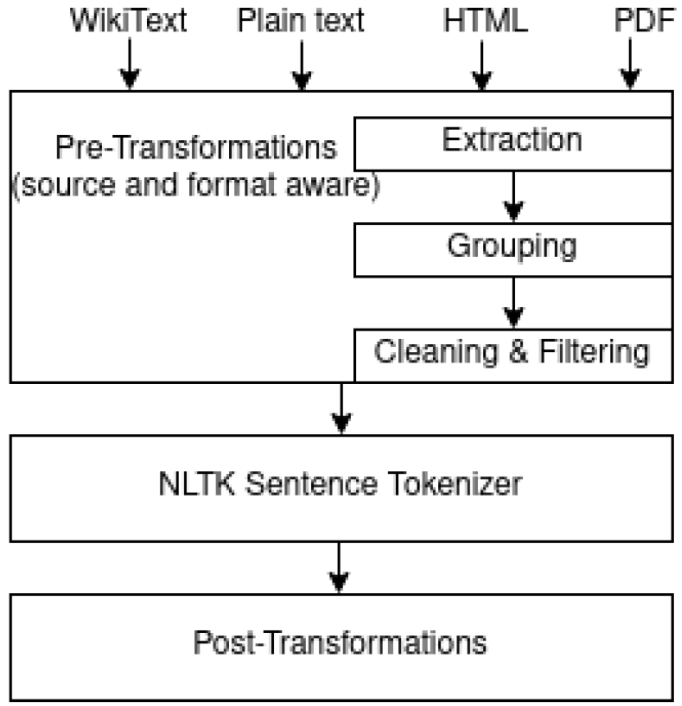
Implementation

NSOAMT

Ingestion pipeline

The high level algorithm for ingestion is:

1. Format conversion to (HTML, WikiText, PDF, etc.) to plain text (or plain text groups).
2. Split text into sentences. Apply sentence transformation procedures (such as hash calculation).
3. Insert the whole document into the database.
4. For each sentence (by order of occurrence in the document):
   4.1. search if the sentence already exists in the database:
      4.1.1. if yes, associate existing sentence to current document.
      4.1.2. if no, insert the new sentence into the database, and then associate to the current document.
5. (Post ingestion) Duplicate sentence elimination.



Implementation

NSOAMT

Hashing Algorithm

The choice of the hashing algorithm to be MD5 were based on speed of calculation, and also the fact that implementations in python v3.6 hashlib and PostgreSQL v12 gave identical results.

Electronic document formats

- Plain text UTF-8 encoded text documents
- WikiText
- Hyper Text Markup Language (HTML)
- Portable Document Format (PDF)

Sentence tokenizer

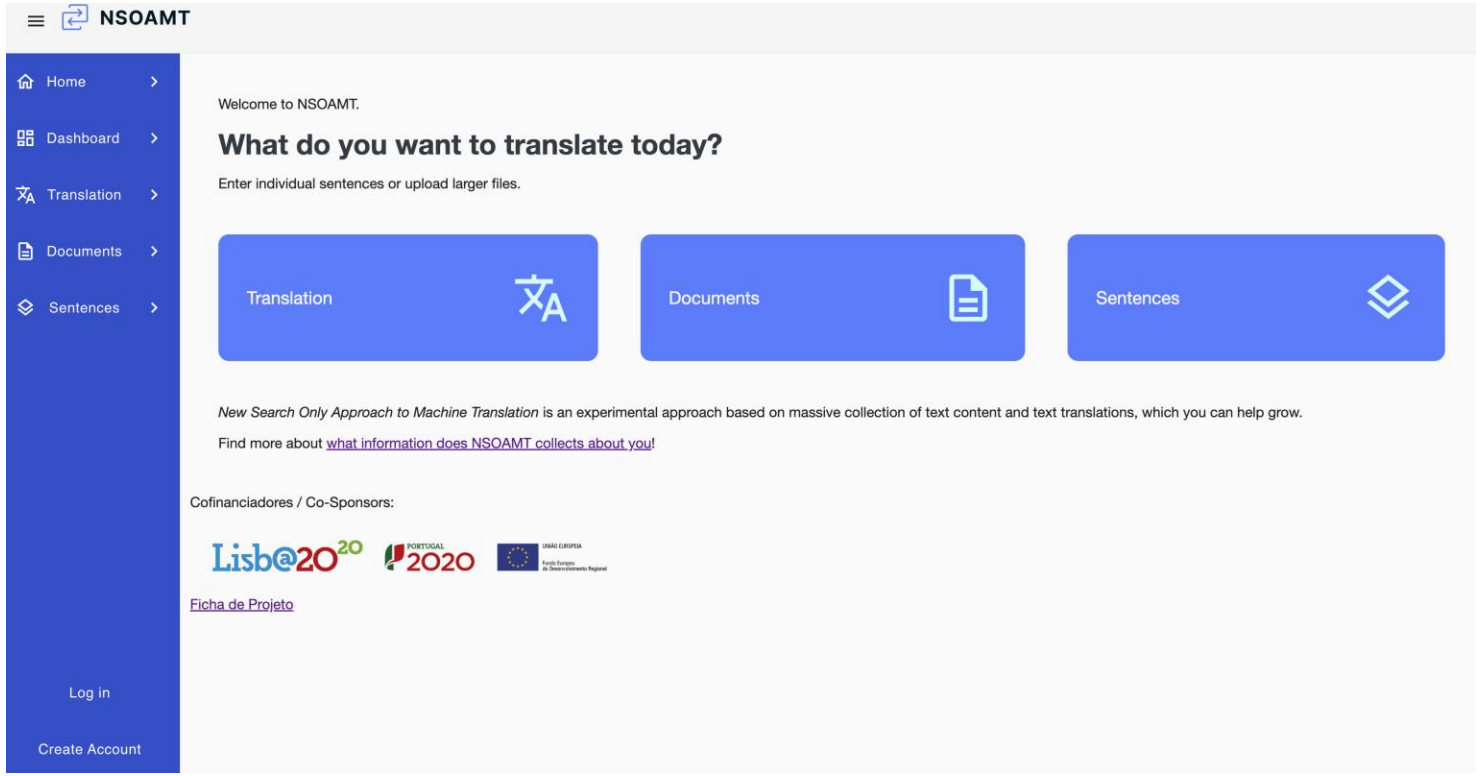
NLTK is a python framework for natural language processing. The default sentence tokenizer API was used to extract an ordered list of sentences from plain text content. It is the core of the ingestion pipeline.

Implementation

NSOAMT

Web Interface

In order to share the knowledge developed and show the planned translation functionality, a web interface for translation based on the technology studied was developed and made publicly available. This site is available at https://nsoamt.pdmfc.com.



DEMO

Latest Results

NSOAMT

Volume of portuguese text ingested from static sources

	Wikipedia (pt)	EUR-LEX (pt)
#documents	8	16,864
#text characters (UTF-8)	1,932,914,020	8,859,532,891
#sentences	16,594,472	34,280,621
(same source)		
#distinct sentences	14,805,146	9,060,610
#distinct sentences %	89.22%	26.43%
(same source)		
#d.sentences with repetitions	255,735	2,070,827
#d.sentences with repetitions %	1.73%	22.86%
(same source)		
#unique d.sentences	14,549,411	6,989,783
#unique d.sentences %	98.27%	77.14%
(same source)		

Latest Results

NSOAMT

Volume of text ingest

	English (en)	Portuguese (pt)
#documents	1,580,741	16 879
#plain text characters (UTF-8)	123,000,703,896	10,794,271,435
	≈114.6 GigaBytes	≈10.1 GigaBytes
#sentences	1,150,174,091	50,884,157
#distinct sentences	885,946,972	23,862,499

Volume of english text ingested from static sources

	ArXiv 0802-2112	Wikipedia (en)	EUR-LEX (en)	tBooks (en)
#documents	1,511,891	61	17,190	51,593
#text characters (UTF-8)	80,399,442,210	14,102,084,349	8,615,499,262	19,883,677,356
#sentences	761,978,703	130,111,846	24,109,494	233,973,998
(same source)				
#distinct sentences	557,332,655	114,423,765	8,112,606	206,490,528
#distinct sentences %	73.14%	85.29%	33.65%	88.25%
(same source)				
#d.sentences with repetitions	18,914,498	2,426,673	1,712,047	5,471,817
#d.sentences with repetitions %	3.39%	2.12%	21.10%	2.65%
(same source)				
#unique d.sentences	538,418,157	111,997,092	6 400,559	201,018,711
#unique d.sentences %	96.61%	97.88%	78.90%	97.35%
(same source)				

Latest Results

NSOAMT

Common #distinct sentences between english sources

Common #distinct sentences (en)	arXiv 0802-2112	Wikipedia (en)	EUR-LEX (en)	tBooks
arXiv 0802-2112	761,978,703			
Wikipedia (en)	46,531	130,111,846		
EUR-LEX (en)	5,448	28,130	24,109,494	
tBooks	63,747	145,199	4,665	233,973,998

Common #distinct sentences (all sources)	2,873

Common #distinct sentences between portuguese sources

Common #distinct sentences (pt)	Wikipedia (pt)	EUR-LEX (pt)
Wikipedia (pt)	16,594,472	
EUR-LEX (en)	8,600	34,280,621

Common #distinct sentences (all sources)	8,600

Next Steps and Challenges

NSOAMT

Text volume

- Address the question: **"How much text volume is needed to achieve a desired percentage of distinct sentences with repetitions?"**
- Seek partners to obtain new sources for text ingestion.

Improvements

- Apply character-level simplification techniques within a sentence, encompassing the removal of punctuation, digits, date tagging, and the utilization of a custom sentence tokenizer.

Web Interface

- Testing with end users to obtain feedback for usability improvements.
- Improve collaborative revision resources to improve the quality of translations.
- Create a public dashboard to monitor translation metrics.

Partners

- Discover new prospective partners to harness the acquired knowledge and enable the creation of innovative solutions/business models.

Questions

Lisb@20[20]
PORTUGAL 2020
UNIÃO EUROPEIA
Fundos Europeus Estruturais e de Investimento