ComPair

Compare documents in any format, in any language

Lingua et Machina

Corporate Multilingual Content Management

September 2018

Presentation

- ComPair is a universal document comparer, to be installed on your computer:
  - It compares two versions of a document,
  - If needed in two different file formats,
  - In the same language or in two different languages.
- All common office file formats are accepted, including scanned documents.
- All languages are processed.

Pricing

- The first month of use is free for any language(s), please feel free to test.
- Thereafter, the cost is 5 € (excl. tax) per language and per month.
- The user can select any number of languages, starting with one.
- Selection of language(s) is done during the license request procedure.
- New language(s) can be added at any time.

Installation

- Download the software here:

  http://lingua-et-machina.eu/ComPair-setup.exe

- Install it on your computer,
- Fill the free license request, send it, receive the license, install it.
- You are ready to start, the following slides show the results for:
  - A comparison of two versions of a document,
  - A comparison of a document and its translation.

Comparing versions

The translation community has seen a major change over the last five years:	0	The translation community has seen a major change over the last five years.
machine translation has become good enough so that it has become advantageous for translators to post-edit it rather than translate from scratch.	1	This is explained by recent developments in statistical machine translation, that is to say the training of a translation engine from a body of existing translated texts.
This is due to recent progress in statistical machine translation, that is, the training of a translation engine with a corpus of existing translations.	2	Thanks to progress in the training of statistical machine translation engines on corpora of existing translations, machine translation has become good enough so that it has become advantageous for translators to post-edit machine outputs rather than translate from scratch.
Current enhancement of machine translation (MT) systems from human post-edition (PE) of raw outputs are somewhat efficient yet rather basic:	3	However, current enhancement of machine translation (MT) systems from human post-edition (PE) are rather basic:
the post-edited output is added to the training corpus and the translation model and language model are re-trained, with no clear view of how much has been improved and how much is left to be improved.	4	the post-edited output is added to the training corpus and the translation model and language model are re-trained, with no clear view of how much has been improved and how much is left to be improved.
	5	Moreover, the final PE result is the only feedback used:
In this approach, only the final PE result is used, no other user feedback on the raw MT quality is provided, such as the cognitive processes of the post-editor or the logging of the post-edition actions he has performed.	6	available technologies do not take advantages of logged sequences of post-edition actions, which inform on the cognitive processes of the post-editor.
The KEHATH project intends to address these issues in two ways:	7	The KEHATH project intends to address these issues in two ways.
Firstly, leverage advanced machine learning (ML) techniques in the MT+PE loop.	8	Firstly, we will optimise advanced machine learning techniques in the MT+PE loop.
Our goal is to boost the impact of PE, that is, reach the same performance with less PE or better performance with the same amount of PE.	9	Our goal is to boost the impact of PE, that is, reach the same performance with less PE or better performance with the same amount of PE.
In other words, we want to improve machine translation learning curves.	10	In other words, we want to improve machine translation learning curves.
For this purpose, active learning and reinforcement learning techniques will be proposed and evaluated.	11	For this purpose, active learning and reinforcement learning techniques will be proposed and evaluated.
In the industrial context of KEHATH, we will have to face challenges such as MT systems heterogeneity (statistical and/or rule-based), and ML algorithms scalability to improve a domain-specific MT.	12	Along with this, we will have to face challenges such as MT systems heterogeneity (statistical and/or rule-based), and ML scalability so as to improve domain-specific MT.
Secondly, quality prediction (QP) on MT outputs is crucial for translation project managers.	13	
We have developped over the years a number of confidence estimation and error detection techniques in the laboratory and we will implement and evaluate them in real-world conditions.	14	Secondly, since quality prediction (QP) on MT outputs is crucial for translation project managers, we will implement and evaluate in real-world conditions several confidence estimation and error detection techniques previously developed at a laboratory scale.
A shared concern will be to work on continuous domain-specific data flows to improve both MT and the performance of indicators for quality prediction.	15	A shared concern will be to work on continuous domain-specific data flows to improve both MT and the performance of indicators for quality prediction.

Comparing translations

FAQ

- I received printed documents, how do I compare them?
  - Download and install the free software tesseract-ocr:
  - https://digi.bib.uni-mannheim.de/tesseract/tesseract-ocr-setup-3.05.02-20180621.exe
  - Scan the document(s) and save as PDF file(s),
  - Ask ComPair to process the scan(s), it calls tesseract-ocr automatically.
- Can I export the table in Word or Excel?
  - No, but there is an icon for copying the table to your clipboard,
  - You can then paste it in a Word or Excel file.
- Are there other versions of ComPair?
  - Yes, it is available online in SaaS mode (Software as a Service),
  - Or we can install it as a dedicated server for a company.

FAQ: Translator’s corner

- I’m a translator, how do I align translated documents?
  - Just select the languages and launch ComPair with your documents,
  - Then click on the “TMX” icon to export it as a standard memory file.
- Can I edit the alignments?
  - Yes, use is the command icon left of the segment number shown in slide 6.
- Can I merge several alignments in one memory?
  - Yes, please call us.

More info?

- Contact : François Brown de Colstoun
  - +33 6 80 95 94 39
  - Skype : f_brown_de_colstoun
  - fbc@lingua-et-machina.com

- http://www.lingua-et-machina.com/