University of Southampton

Digital Humanities

Digital Humanities: Exploring Natural Language Processing Applications with Modern Language And Linguistics

Isobel Lester, 2022

Contents

Supplementary Resources

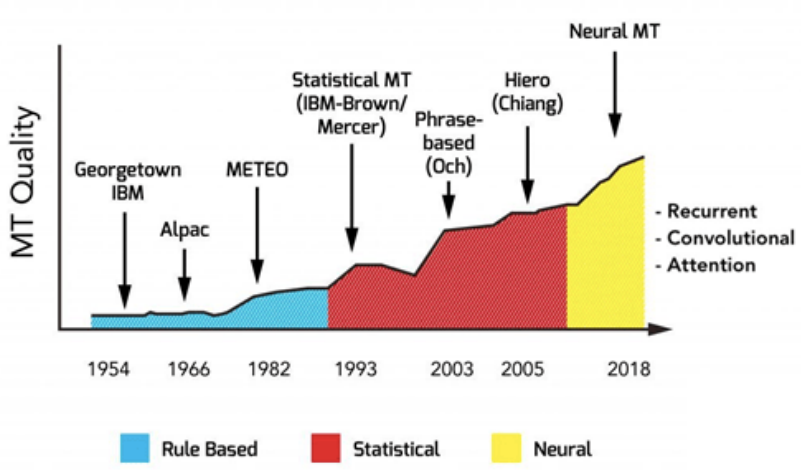
Github Repository (link)

This project was undertaken in Spring/Summer 2022 by Isobel Lester as part of a Digital Humanities internship funded by the School of Humanities at the University of Southampton

Overview

Within digital humanities, we aim to reframe questions within the humanities through the lens of digital and computational methods. One such method is natural language processing, a branch of artificial intelligence which strives to give computers the ability to understand written and spoken human texts similarly to how humans do. This offers modern language and linguistics researchers enhanced abilities to analyse large datasets at scale. Therefore, this project aims to explore the applications of natural language processing within the humanities, in particular machine translation and sentiment analysis. It compares corpora assembled from human and machine translations of text in order to create insights into these applications' merits and limitations, and use cases for integrating computation methods into traditional humanities research.

A history of machine translation



https://syncedreview.com/2020/05/20/neural-network-ai-is-the-future-of-the-translation-industry/

The first machine translation was produced in January 1954 by an electronic 'brain' in Georgetown University and IBM's collaborative research in computation. In this transmission, 60 brief statements about politics, law and military affairs among other topics were submitted in Russian to the 701 computer, which within a few seconds produced sentences of easily readable English. This breakthrough came after around a decade of researching how to make a machine that would convert the meaning of words clearly from one language to another. No results were achieved until Georgetown enlisted the aid of the IBM 701, the most versatile electronic 'brain' to date.

To achieve this, the electronic translators created an entirely new electronic language, where they take words and attach tags or signs to them that denote rules of grammar and meaning which gives greater precision. Translation was only possible because of these rule-

tags attached to normal words, which acted as instructions prepared by humans, since the IBM 'brain' could not think independently. These rule tags govern phrases where transposition of words was required to make sense, choice of meaning when a word has several definitions, omissions of words that are not required in the target text, and insertion of words needed for it to make sense in the target text.[1]

Statistical MT

Following from the rule-based machine translation model of 1954, Brown and Mercer of IBM's research division published 'A Statistical Approach to Language Translation'[2] which outlined the next key framework of machine translation in 1988. Statistical machine translations learn to translate by exploring extensive human translations know as bilingual corpora in order to create statistical patterns on the likelihood that a given sentence in the source language corresponds to a sentence in the target language[3]. Additionally, the statistical model was phrased-based, thereby increasing accuracy compared to rule-based models which worked word by word and thus lost context of the rest of the phrase.

Translation Today: Neural Machine Translation

It was the introduction of neural networks in 2014 that really took machine language learning to the next level and have generated the software models we see today. Instead of computers relying on an analysis of human translators to produce rule-based translations, neural networks entail a machine learning algorithm whereby users train the translation machines to recognise source and target connections from extremely large training datasets. One key difference from rule-based translations is that users don't necessarily tell the deep learning algorithm exactly what to look for, rather the algorithm is left to explore the data and observe correlations itself. From this, the algorithm predicts the likelihood of certain sequences of words and produces accordingly.[3] These neural networks are modelled on the human brain and human learning process. Like the human brain when learning, when you make a mistake, you back up and keep trying again until you understand. This means that neural machine translators learn and adapt as they encounter new data, leading to more accurate translations over time. Neural machine translation is the standard model for translation software today, including Google Translate which made its full transition to neural networks in September 2016 due to the fewer engineering and design choices required and increased speed and accuracy.

---

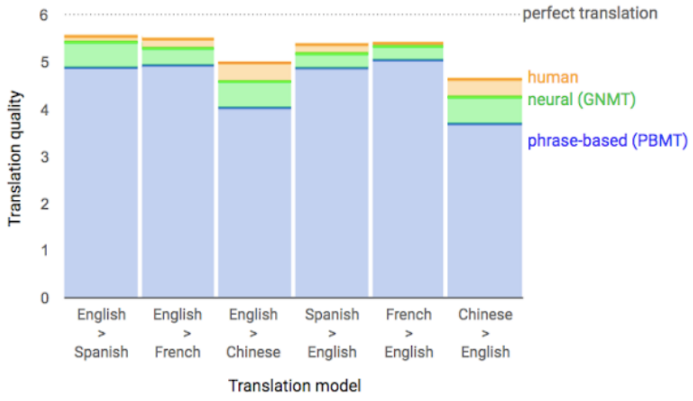
[1] IBM (1954). " 701 Computer". IBM Press Release. Available: https://www.ibm.com/ibm/history/exhibits/701/701_translator.html

[2] Brown & Mercer, 1988. "A Statistical Approach To Language Translation". IBM Research Division. Available: https://dl.acm.org/doi/pdf/10.3115/991635.991651

[3] Žďárek, 2021. "Machines that Think: The Rise of Neural Machine Translation. Available: https://www.memsource.com/blog/neural-machine-translation/

With neural machine translations offering increased accuracy and quality whilst requiring little supervision, many often wonder if AI translation will render human translators obsolete. However, the US Bureau of Labor Statistics actually predicts a 24% grown for employment of interpreters and translators between 2020 and 2030 which is well about the average growth rate of 8% [4], meaning that human translators will still be in demand since there are still significant limitations to neural models such as:

- Difficulties with processing cultural references, idiomatic expressions etc.
- Difficulties translating legal or technical information, particularly when abbreviations are used.
- The model's architecture only enables sentence-like length sequences to be processed at one time meaning that when longer sequences are inputted the model processes sentence by sentence without acknowledging preceding sentences. This means that they can produce inconsistencies throughout the text as they lose context.
- For translating between some languages there is not as much training data available which will impact on accuracy.



https://ai.googleblog.com/2016/09/a-neural-network-for-machine.html

Therefore, this is not to say that machine translations will outperform and make redundant human translators but instead will cause the industry to evolve to an integration between human and computer translation. With the use of these translators only set to grow, it is important to understand their merits, limitations, and capacities compared to humans in order to implement them most effectively which this project aims to explore.

<u>Key questions</u>

- How accurately can a machine translate emotion and the desired impact of the author compared to a human translation?
- What variations are produced between a human generated and AI generated translation? Do they alter the meaning?
- What limitations are there to AI translation software?
- Should complex human languages be reduced to numbers? How accurately can it do this?

[4] https://www.bls.gov/ooh/media-and-communication/interpreters-and-translators.htm

Methodology

Collating Data

In order to see how the machine translator and sentiment analyser operated a variety of contexts, I collected data from several sources including: press releases, instruction manuals, fictional books, TV subtitles, and website articles.

Performing Sentiment Analysis

Alongside machine translation, another application of NLP is sentiment analysis which attempts to extract subjective qualities from text and quantify its emotional intensity.[5] As part of this project, I have created a Juptyer Notebook which uses the Sentiment Analyzer function from a widely used NLP toolkit – NLTK – on the data files mentioned above.

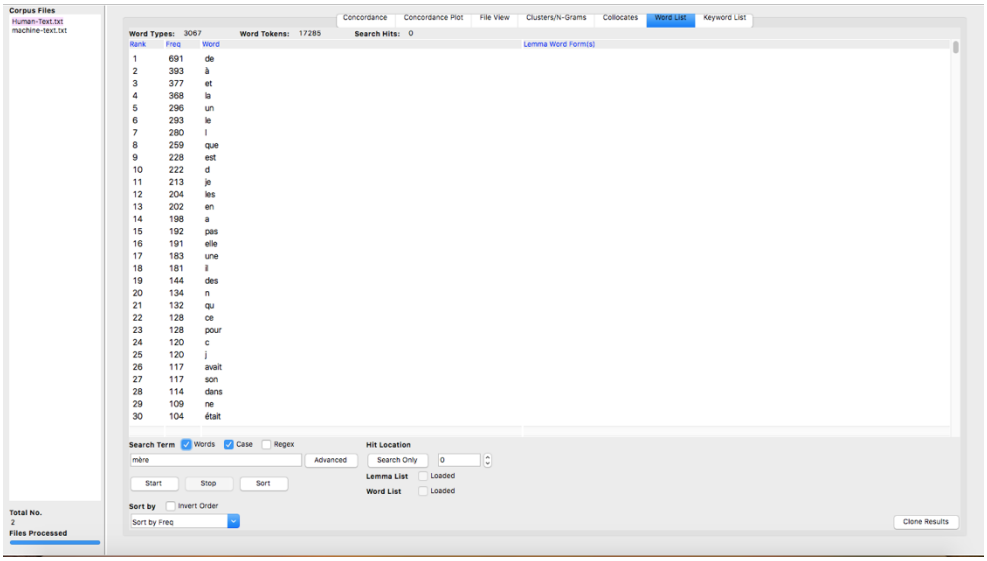
To see this output and interactive notebook are available here.

Antconc Analysis

For qualitative analysis, I used AntConc, an advanced corpus analysis application which provides details about texts inside text files. The main features I used where:

Word List

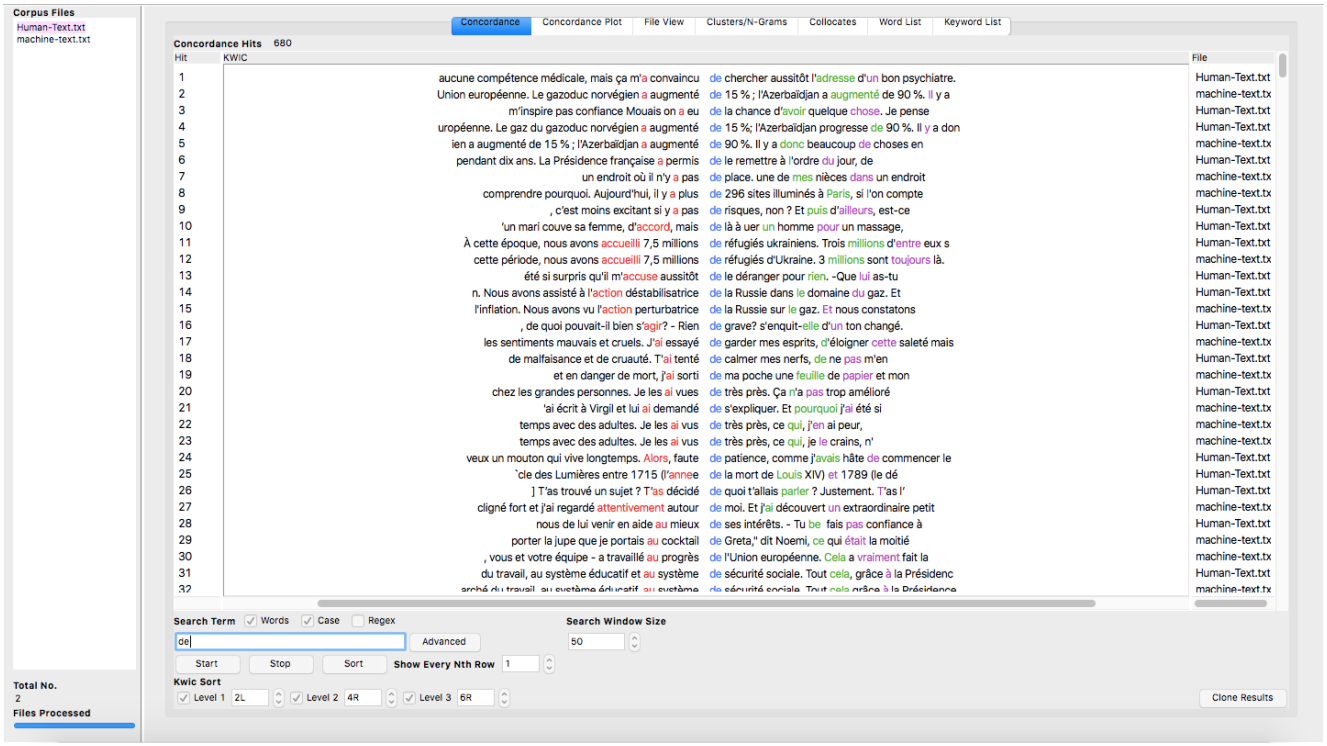
Word list is a good place to start when analysing a new corpus. It highlights the most common words in a corpus. By clicking on each word, you see the Keyword in Context (KWIC) and can start looking at patterns and significant features of the corpus.



[5] IBM Cloud Education (2020). "Natural Language Processing (NLP). Available: https://www.ibm.com/cloud/learn/natural-language-processing?utm_medium=OSocial&utm_source=Youtube&utm_content=000027BD&utm_term=10004432&utm_id=YTDescription-101-What-is-NLP-LH-Natural-Language-Processing-Guide

Concordance

The concordance tool enables you to search terms in context.



Observations

Analysing Machine Translations

This first observation section aims to highlight differences produced between human and machine translations, in particular highlighting any variations that create differences in meaning between the source text and the target texts, any differences in tone and structure, and finally how the machine handles cultural information or idiomatic expressions.

When looking at these examples, I would like to point out these are observations from a selection of phrases out of thousands of such phrases in the data. A characteristic of the qualitative approach to data analysis is that it is difficult to make generalisations and conclusions based off a few examples, so whilst the below examples highlight interesting observations in the data there may also be examples in the data which illustrate an opposing point. Therefore, these should be taken as a collection of observations rather than as definitive conclusions.

Semantic Variation

« La mère de Catalina était morte à son tour **deux ans** plus tard » (Human)

« La mère de Catalina est décédée **quelques années** plus tard » (machine)

This extract highlights how a translator's interpretation can create a different meaning. In the original English text, it says "a couple of years" which the human translator has taken to mean two, and the machine an unspecified number.

« Catalina avait employé une écriture tremblante, peu soignée » (human)

« L'écriture semblait instable, bâclée » (machine)

```
message_textMT = "L'écriture semblait instable, bâclée "

message_textHT = "Catalina avait employé une écriture tremblante, peu soignée."

scoresMT = SIA.polarity_scores(message_textMT)

scoresMT
```

```
{'neg': 0.385, 'neu': 0.615, 'pos': 0.0, 'compound': -0.3612}
```

```
scoresHT = SIA.polarity_scores(message_textHT)

scoresHT
```

```
{'neg': 0.2, 'neu': 0.6, 'pos': 0.201, 'compound': 0.0018}
```

In the human example, “Catalina” is the subject who is writing in a shaky manner, whereas in the machine example, there is not an agent and instead emphasis is put on the object (the writing). Whilst the machine translation ultimately creates the same picture of shaky or messy handwriting, it’s interesting that when subjected to sentiment analysis this sentence produces a more negative result than the human translated text.

This difference may be down to the fact the sentiment analyser incorrectly detected “soignée” – meaning “carefully” – as a positive statement and has not considered it to be connected to the “peu” preceding it, which changes the meaning of the passage to “with little care” or “carelessly”. A human reader would likely read this as a negative statement.

Register

« **Veuillez visiter** le site Web de Samsung » (Human)

« **Visitez** le site Web de Samsung » (Machine)

Both extracts give instructions to reader. However the human translated example has used “veuillez”, which is generally a more eloquent and polite way of suggesting to the reader to fuflfil an action. Here then the human translation has adopted a more formal tone than the machine translation.

« **Vous avez** la lettre ? » « Oui, la voici. » (machine)

« **Tu as** encore la lettre? » « Oui. La voici. » (human)

« Tu **ne fais pas** confiance à Virgil » (machine)

« **Tu fais pas** confiance à Virgil? » (human)

These extracts were highlighted as they illustrate a commonly cited limitation of neural translation: understanding tone and consistency throughout the text. In French “vous” is reserved for formal conversations whereas “tu” is used in informal conversations. These extracts are taken from a casual conversation so “tu” would be most appropriate, but the machine translation has deemed the formal “vous” more appropriate. Additionally, in spoken French it is common to omit the “ne” from a “ne … pas” negative construction. The second human extract features that, but the machine does not because – we can assume – machine translation is not well equipped for handling and generation colloquial speech. Finally, a limitation of neural translation is that it works on a sentence-by-sentence basis and does not factor in preceding sentences, which in this instance has led to inconsistency with the use of “tu” and “vous”, with the latter being used in the first example and “tu” in the prior despite the fact both sentences are spoken by the same character.

Technical language

« Connexion au MDC grâce à **un adaptateur RS232C**. » (human)

« Se connecte à plusieurs produits à l'aide **d'un câble série.** » (machine)

Technical language, particularly when it involves abbreviations, can pose a challenge for machine translators. This extract originates from an instruction manual of an electronic good. Here, we can see that the human translator references a specific adapter whereas the machine has failed to interpret the exact adapter and instead made a broader statement. In certain contexts, the lack of specificity may enable the translation to work, however in the case of an instruction manual it will likely reduce the quality of the translation.

Idiomatic Expressions

« Je pense que Buddy Holly **pédale dans le yaourt**. » (human)

« Buddy Holly **n'a pas l'air d'être un très bon serveur** » (machine)

« L'argument **fit mouche.** » (human)

« Il **avait raison** » (machine)

Idiomatic expressions are common in all languages and suggest a higher fluency of the language since they do not necessarily make logical sense. In the first extract, the human translator has used “pédale dans le yaourt”, translating literally to “pedal in yoghurt” an idiomatic phrase that means to not make progress or not get anywhere. In the second extract the human translator has used “faire mouche” which means to hit the nail on the head. In both extracts for the machine was given a sentence in non-idiomatic English which it has successfully translated into meaningful into French. However, these examples illustrate that whilst a machine can get across general meaning, it doesn’t possess a similar aptitude with the language as a human.

Anglicism

« Garanti pour être **un line-up** diversifié chaque année » (machine)

« Ce festival garantit **une programmation** différente chaque année » (human)

« Déplacer le curseur **sur Eteindre** » (Human)

« Déplacer le curseur **sur Power off** » (machine)

Due to globalisation and other cultural-linguistic exchanges, English has become increasingly common in colloquial spoken French. In these examples, it is interesting that the machine translation uses the English terms “line-up” and “Power off” since the use of English is informal and does not follow specific rules. Instead, this likely means the machine has failed to comprehend the meanings of these words and so decided to keep them in English.

Cultural References

« d’affiche les **Red Hot Chili Peppers**, **The** Weeknd, **Imagine Dragons**, Lana Del Rey et bien » (human)
« d'affiche comme **Red Hot Chili Peppers, The** Weeknd, **Imagine Dragons**, Lana Del Rey, et bien » (Machine)

A common critique of machine translations is that they struggle with cultural references, however, in this example the machine translation has successfully identified the band and artist names and left them in their original form.

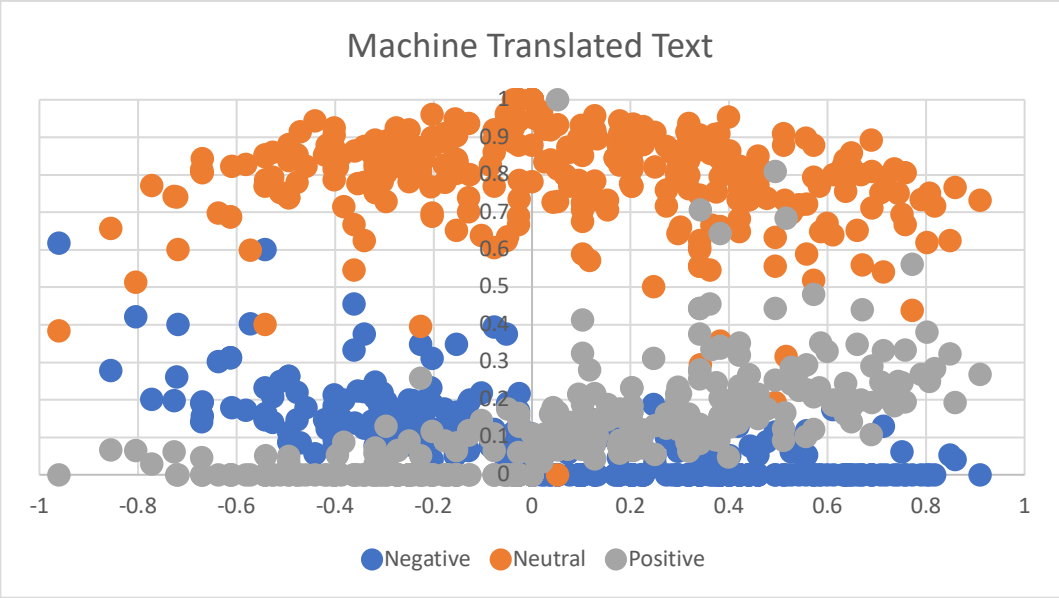
« Donc le couple quittant la fête **à dix heures du soir** a donc a brisé la convention. » (machine)

« Aussi, lorsqu'un couple s'en échappa **à 22 heures**, la convention implicite vola en éclats. ». (human)

This extract highlights an instance where the machine translation has produced lower quality translation than the human translator. In English to express 10PM, you could write it as 10PM or 10 in the evening based off the 12 hour clock, but in France, the 24 hour clock is used, so to express 10PM “22 heures” would be most accurate.

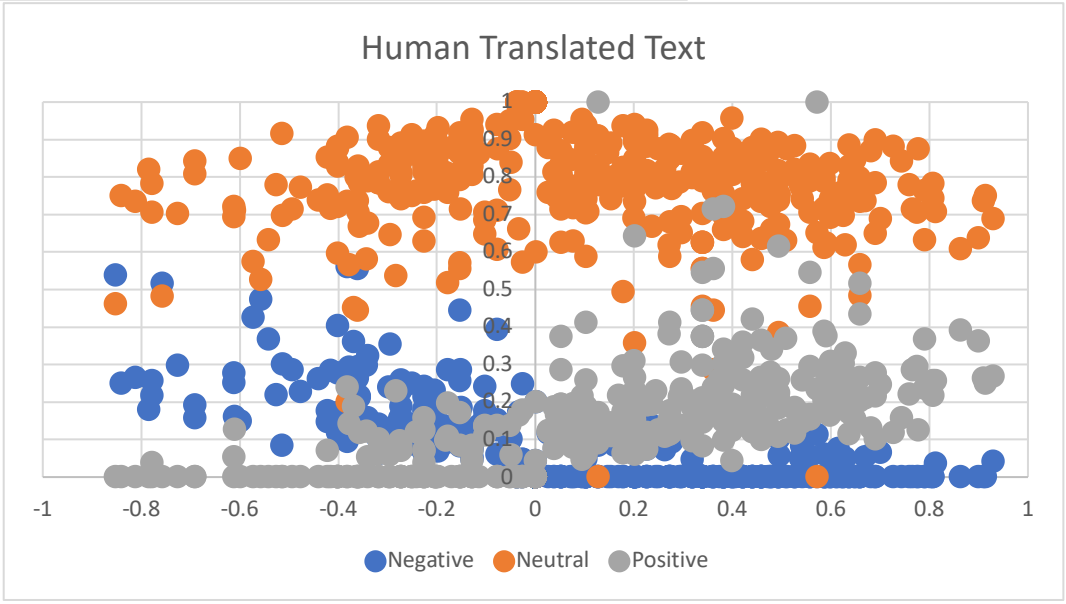
Exploring Sentiment Analysis

	Compound	Negative	Neutral	Positive
Machine	0.050899	0.050971	0.868634	0.07857
Human	0.086132	0.052433	0.845264	0.100393



X Axis: Compound score. A sum of positive, negative and neutral scores which is placed on a scale between -1 (most negative) to +1 (most positive).

Y Axis: Negative score (blue), neutral score (orange), positive score (grey)



Whilst subtle, the averages across the corpus of Human Translated Texts (“human_data”) and Machine Translated Texts (“machine_data”) suggests that human translations tend to carry more emotion than the machine translations, and in particular that humans translations are more positive than machine translations.

Negation

« **qu'elle n'avait pas envie d**'écrire beaucoup » (machine)

«**ne prenait pas le temps d**'écrire » (human)

```
message_textMT = " elle n'avait pas envie d'écrire beaucoup "
```

```
message_textHT = "elle ne prenait pas le temps d'écrire. "
```

```
scoresMT = SIA.polarity_scores(message_textMT)
```

```
scoresMT
```

```
{'neg': 0.0, 'neu': 0.867, 'pos': 0.133, 'compound': 0.0191}
```

```
scoresHT = SIA.polarity_scores(message_textHT)
```

```
scoresHT
```

```
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
```

This extract outlines how the sentiment analyser can sometimes create results that a human reader would not necessarily agree with. Here the sentiment analyser has interpreted the machine translation as have a slightly positive tone (0.133). However the sentence “she didn’t want to write” is a negation expressing a lack of desire to fulfil an action and therefore one which a human would consider as emotionally negative.

« Catalina **privilégiait** la machine à écrire » (machine)

« Catalina **préférait** la machine à écrire » (human)

```
message_textMT = "Catalina privilégiait la machine à écrire"
```

```
message_textHT = "Catalina préférait la machine à écrire "
```

```
scoresMT = SIA.polarity_scores(message_textMT)
```

```
scoresMT
```

```
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
```

```
scoresHT = SIA.polarity_scores(message_textHT)
```

```
scoresHT
```

```
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
```

This extract is interesting since the sentiment analyser has failed to detect any positive or negative sentiment in the sentence. A human reader is, however, likely to detect a slight positive sentiment.

« Les fêtes chez les Tuñons se terminaient toujours **très** tard » (machine)

« Les fêtes chez les Tuñon se terminaient toujours **affreusement** tard » (human)

```
message_textMT = "Les fêtes chez les Tuñons se terminaient toujours très tard "

message_textHT = "Les fêtes chez les Tuñon se terminaient toujours affreusement tard"

scoresMT = SIA.polarity_scores(message_textMT)

scoresMT
{'neg': 0.0, 'neu': 0.789, 'pos': 0.211, 'compound': 0.34}

scoresHT = SIA.polarity_scores(message_textHT)

scoresHT
{'neg': 0.218, 'neu': 0.602, 'pos': 0.18, 'compound': -0.128}
```

In this extract, the machine and human translator has used different adjectives (very late vs. awfully later) that portray the same meaning. The sentiment analyser has read the word 'awfully' and prescribed a negative sentiment that a human translator – who would understand the context of this phrase – would not.

Additionally, this extract also illustrates the reduced accuracy of the machine translators, since when referring to a family in French it is correct to express their name in the singular form: so "les Tuñon" rather than "les Tuñons".

This pattern is also seen in the below extract where both convey the meaning that it was a very short engagement, but the use of "ridiculously" has meant that that the human translation has been assigned a slightly negative sentiment. Since both these extracts overall convey the same meaning, we can infer that it is the different adjective choice that has created a variation in the sentiment analysis, and therefore that a sentiment analyser may not particularly accurate in interpreting a whole phrase, and is susceptible to influence from small changes.

```
message_textMT = "Les fiançailles de Catalina avaient été très courtes "

message_textHT = "Les fiançailles de Catalina avaient été ridiculement courtes "

scoresMT = SIA.polarity_scores(message_textMT)

scoresMT
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}

scoresHT = SIA.polarity_scores(message_textHT)

scoresHT
{'neg': 0.136, 'neu': 0.864, 'pos': 0.0, 'compound': -0.0258}
```

« Je pense que Buddy Holly **pédale dans le yaourt**. » (human)

« Buddy Holly **n'a pas l'air d'être un très bon serveur** » (machine)

```
message_textMT = "Buddy Holly n'a pas l'air d'être un très bon serveur "
```

```
message_textHT = "Je pense que Budy Holly pédale dans le yaourt.    "
```

```
scoresMT = SIA.polarity_scores(message_textMT)
```

```
scoresMT
```

```
{'neg': 0.0, 'neu': 0.806, 'pos': 0.194, 'compound': 0.4391}
```

```
scoresHT = SIA.polarity_scores(message_textHT)
```

```
scoresHT
```

```
{'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
```

This extract outlines a limitation of sentiment analysis. The human translator has used an idiomatic expression whereas the machine has given a literal translation. However, the sentiment analyser has failed to understand the meaning of the idiom.

From looking at the graphs at the top of this section, it appears that there are more neutral data points among the human data than the machine data, which contradicts the findings that machines tend to be more neutral and human translators more emotional. This example illustrates that this may be due to the fact that the sentiment analyser is more likely to fail to understand the sentiment behind human translations with more complex use of language.