MT, AI and the Language Professional

How to fit in to the emerging language services landscape

Jay Marciano
Director, Machine Translation Outreach & Strategy
Lengoo GmbH
jay.marciano@lengoo.com
Introductions

Jay Marciano

• Director of Machine Translation, Lengoo
• 25 years’ experience in the development & application of MT
• President of AMTA (Association of Machine Translation in the Americas)
• Avowed language geek on a mission to increase understanding, cooperation, and collaboration among translators and interpreters, MT researchers and developers, language service companies, and customers

Lengoo

• Founded 2014
• Headquartered in Berlin
• Language Technology and Services Provider
• ~100 employees
One thing has become very, very clear to us in the last few years ...

Technology is advancing faster and faster
In 2003, Ray Kurzweil estimated that the rate of technological advance doubles every decade.

Assuming that’s true, it is exponential growth, with the technological advance of the past decade helping to increase the rate of development in the next decade.

In 2023, we’re developing technology twice as fast as we were in 2013, and 128 times faster than in the 1950s, when MT was first demonstrated.

And from now until 2033, the rate of change will double again.
Drivers of Progress

Transforming **Material**
- Pre-industrial Age
- Stone Tools
- Bronze Tools
- Iron Tools

Transforming **Energy**
- Industrial Age
- Wind Power and Hydropower
- Steam Power
- Electricity
- Internal Combustion

Transforming **Information**
- Information Age
- Computational Power
- Information as Building Blocks

Timeline:
- 2,000,000 BCE
- 3,300 BCE
- 1,200 BCE
- 100 BCE
- 1848
- 1895
- 1930
- 1973
- 2009

Copyright 2023 by Jay Marciano
In the Information Age, data is the primary driver of progress.
And it’s an abundant resource!
Let’s think about Complexity
Life is complicated
The intersection of simplicity and endless variability
How many ways can you shuffle a deck of cards?
How many ways can you shuffle a deck of cards?

\[ x = 52 \times 51 \times 50 \times 49 \times 48 \ldots \times 3 \times 2 \times 1 \]

\[ x = 52! \]
How many ways can you shuffle a deck of cards?

Would it surprise you to hear that there are

824 TRILLION

ways to shuffle a deck of cards?
How many ways can you shuffle a deck of cards?

And what if that 824 trillion were only this much of the true number?

80,658,175,170,943,878,571,660,636,856,403,766,975,289,505,440,883,277,824,000,000,000,000
You think that’s a big number?
Language is unimaginably more complicated!

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>A Deck of Cards</th>
<th>A Human Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of “words”</td>
<td>52</td>
<td>Let’s say 100,000</td>
</tr>
<tr>
<td>Possible lengths</td>
<td>1</td>
<td>Let’s say 100</td>
</tr>
<tr>
<td>Repetition of “words”?</td>
<td>Impossible</td>
<td>Possible</td>
</tr>
<tr>
<td>Possible combinations</td>
<td>$8.0658175 \times 10^{67}$</td>
<td>Incalculably higher</td>
</tr>
</tbody>
</table>

And yet you’re able to understand multiple languages.
Life is complicated

• You have a million things spinning around in your head.
• But you need to GET STUFF DONE.
• So you calm down and decide to attack a problem.
• You focus on the task at hand and say to yourself, “Take it one step at a time ... just figure out what to do next and you’ll get there eventually.”
You’re already thinking algorithmically

**algorithm n.** a procedure or formula for solving a problem, based on conducting a sequence of specified actions
Algorithmic task completion

Determine next task → Collect information → Determine best next step → Perform step
Is task completed? → Yes
→ No

Copyright 2023 by Jay Marciano
Algorithmic job completion

1. Get new job
2. Determine next task
3. Collect information
4. Determine best next step
5. Perform step

Is job completed? (Yes/No)

Is task completed? (Yes/No)
Algorithmic job selection and completion

1. Get new job
2. Is job worth doing?
   - Yes: Determine next task
   - No: Is job completed?
     - Yes: Collect information
     - No: Determine best next step

3. Collect information
4. Determine best next step
5. Perform step

6. Is task completed?
   - Yes: Is job completed?
     - Yes: Complete
     - No: Go back to get new job
   - No: Go back to determine next task
Algorithmic goal selection and completion

1. Get new goal
2. Is goal worth pursuing?
   - Yes: Get next job
   - No: Has goal been achieved?
     - Yes: Stop
     - No: Get new goal
3. Get next job
4. Is job worth doing?
   - Yes: Determine next task
   - No: Is job completed?
     - Yes: Collect information
     - No: Determine best next step
5. Collect information
6. Determine best next step
7. Perform step
8. Is task completed?
   - Yes: Stop
   - No: Is job completed?
     - Yes: Determine next task
     - No: Is job worth doing?
       - Yes: Collect information
       - No: Determine best next step
9. Determine next task
10. Collect information
11. Determine best next step
12. Perform step

Copyright 2023 by Jay Marciano
And that’s what AI and Neural Networks do ...

- Learn about a task
- Develop a method for:
  - Determining what the best next step is
  - Executing that step
  - And repeating that process until
    - the task is done,
    - or the job is done,
    - or, maybe someday, an overarching goal is achieved.
Let’s think about Machine Translation
Quickening development of Machine Translation

Rules-based MT
Build a translation with language-specific algorithms and bilingual dictionaries
Was state-of-the-art for ~50 years

Statistical MT
Find the most probable translation using statistical modeling
Was state-of-the-art for ~17 years

Neural MT
Predict a translation using a deep neural network
State-of-the-art for ~6 years
Let’s build an NMT model
Monolingual training material for NMT

This is a general principle, which we can state:

More specifically, the presidency currently believes that, should this type of intelligence capability be established within the EU, we would have to investigate the procedures that would need to be implemented in order to ensure democratic monitoring of these activities.

These procedures would naturally have to take into account the particular nature of the intelligence field which, in order to function effectively, requires that information gathered remains confidential.

Mr President, to sum up and in conclusion, I would like to congratulate Mr Watson once again for his excellent work and I hope that you have an interesting debate.

I shall be staying in the House to listen to what you have to say.

Mr President, ladies and gentlemen, recently, when I had to explain the function of our Echelon Committee, I used the following example: I said, 'Imagine you are a detective sergeant.

In short, it is a rather difficult case.

This is tantamount to what happened when journalists began to tell us that there was a global interception system called Echelon which was operated by the United Kingdom, Canada, Australia and New Zealand and the United States, the United Kingdom, Canada, Australia and New Zealand and the United States, the United Kingdom, Canada, Australia and New Zealand and the United States.

Individuals, businesses and institutions, they said, were subjected to systematic world-wide surveillance.

福建厦门即日起到今年年底将采取临时性限购措施，规定每个购房家庭只能在进行调整。

10月1日，广州市国土房管局对当地在售楼盘进行了检查，要求各开发商在购
房须知上对新一轮房地产调控政策有所体现。目前他们正密切关注楼市动态，
一旦发现房价出现大幅反弹，不排除采取限购等措施，保持市场稳定。

深圳日前出台了房产限购令，规定拥有深圳市户籍的居民家庭，限购两套房，
持有一年以上纳税证明或者社会保险缴纳证明的非户籍居民家庭，限购一套房
处。

各地楼市观望气氛浓厚

热热闹闹的北京楼市在“黄金周”突然遇冷。根据北京市房地产交易管理网的数
据显示，9月30日，因为“末班车”效应，北京二手房签约量出现井喷，达到67套，也是新政后首次突破千套大关。但到了10月1日和10月2日，交易量就
迅速跌至谷底，两天共成交了18套。

在上海，被称为楼市“晴雨表”的“假日楼市?秋季房展”3日上午开幕。尽管这被
认为是近3年来最热闹的一次房展会，尽管部分楼盘优惠幅度巨大，“直降10万
...
What is “learned” about words

For every word in the training material, the Deep Learning system calculates a word embedding, a vector that contains semantic and grammatical information and indicates relationships among the words.

This information provides a multidimensional map of each supported language, showing the relationships among all of the words in those languages.
Word embeddings

- Word embeddings can be mapped in multidimensional space
- Similar words have similar values (or locations)
- The mathematical relationship between words that have a related meaning resembles the relationship between two other words that share that semantic relationship
  - Example: Countries and their capital cities
Yee King Hotel is located at the heart of Mong Kok, Hong Kong’s busiest commercial shopping district. Very convenient public transportation nearby for visiting, dining, and entertainment.

Close to a wide range of jewelry stores, designer clothing shops, and cosmetics shops.

Adjacent to Langham Place, Mongkok electronics district, Ladies Market, Sneaker Street, and a variety of popular eateries, restaurants, as well as Sasha, Bonjour, the Watsons, Mannings, Chow Sang Sang, Chow Tai Fook, Feng Chak, and Broadway.

Comprehensive traveling services for your most wonderful vacation in Hong Kong.

Yee King Hotel is located in Mong Kok at the heart of Hong Kong’s busiest commercial shopping district. Very convenient public transportation, close to Mongkok MTR station, railway station, A21 airport bus station, and Sino-Hong Kong express bus station.

Cozy mini loft close to Wuhe Station at Metro line no.5

Please refer to the photos for the room interior.

I am an interior designer and sell skin care product through micronet as my part-time job. I'm very particular in the quality of living and I love to decorate my living space. I hope you’ll appreciate and cherish my place like I do.
Training an artificial neural network

1. Make prediction using the latest weights and biases.
2. Assess prediction by comparing to reference value.
3. Calculate errors through backpropagation.
4. Adjust neurons by updating weights and biases.

Copyright 2023 by Lengoo GmbH
Mastermind

Try it: https://www.webgamesonline.com/mastermind/
Let’s focus on using MT
How can **Machine Translation** be applied to a project?

Part 1: Simplified view of a workflow with TM but *without* MT
How can **Machine Translation** be applied to a project?

Part 2: Simplified view of a workflow *with* MT and post-editing
Translation Tasks with **TM and MT** (old-school)

1. **Begin Project**
2. **Familiarize yourself with Project**
   - Open first segment
   - Determine Segment Type
   - **MT**
     - Apply 2-second rule
     - **Pass**
       - Read source carefully
       - Read target carefully
       - Upload MT Draft
     - **Fail**
       - Delete MT Draft
       - Read source carefully
       - Translate from scratch
   - **TM**
     - Determine Match Percentage
     - **Exact**
       - Read source
       - Check target for context
       - Make needed changes
     - **Fuzzy**
       - Determine differences in source
       - Read target carefully
       - Make needed changes
3. **Save Segment**
4. **Move project to QA**
5. **Yes**
   - More segments?
5. **No**
6. **Open next segment**

---

Using all available project resources (glossaries, style guides, instructions)
MT Quality: Where were we four years ago?

Edit-distance analysis at large LSP (2019)

- Google (AutoML), Microsoft Custom Translator, DeepL
- All customers
- All languages
- All content

![Edit Distance Analysis (2019)](chart.png)

Copyright 2023 by Jay Marciano
MT Quality: Where are we now?

Edit-distance analysis at Lengoo (Q1 2022)

Lengoo Edit Distance Analysis (Q1 2022)

- Lengoo NMT
- All customers
- All languages
- All content

Copyright 2023 by Jay Marciano
MT Quality: Where are we now?

Edit-distance analysis at Lengoo (Q3 2022)

Lengoo Edit Distance Analysis (Q3 2022)

- Lengoo NMT
- All customers
- All languages
- All content
Conditions that you need to adjust to now

1. MT is still improving
2. Reliable automatic quality estimation will revolutionize translation services
1. Improving MT: Context sensitivity

Information throughout a document informs the translation of every sentence

- MT has traditionally worked on a sentence-by-sentence basis
- Potentially valuable information from preceding or following sentences is not currently leveraged
- This is changing and will help in
  - Resolving pronoun antecedents from previous sentences
  - Reducing inconsistencies
  - Assist in domain resolution for vocabulary choice
1. Improving MT: Context sensitivity

Information throughout a document informs the translation of every sentence.
2. Reliable automatic quality estimation
Reducing the “cognitive load” of post-editing on a high percentage of segments

A neural network trained on LQA data, including:

- Source segments
- Unedited MT output
- Post-edited MT output
- Edit distance and other automated metrics
- Language Quality Assessments where available

That predicts:

- whether the machine translation of a new sentence will be in the group of segments that require no change by the translator
2. Reliable automatic quality estimation

When systems can confidently predict which sentences will fall into the 0% Edit Distance Range, the translation market will change radically.
2. Reliable automatic quality estimation

Part 2: Simplified view of a workflow with MT, QE, and post-editing
Translation Tasks with **TM, MT and QE**

**Begin Project**

- Familiarize yourself with Project

- Open first segment

**TM**

- Determine Segment Type
  - Quality Estimation Routine
  - Determine Match Percentage

- Fail

- Pass
  - Exact
  - Fuzzy

**MT**

- Read source carefully

- Read target carefully

- Make needed changes

- Check target for context

- Make needed changes

**Yes**

**More segments?**

**No**

- Move project to QA

**Save Segment**

- Using all available project resources (glossaries, style guides, instructions)

Copyright 2023 by Jay Marciano
Now let’s move beyond MT
Conditions that you need to adjust to

1. MT is still improving

2. Reliable automatic quality estimation will revolutionize translation services

3. Multimodal “Foundation Models” will provide the next MT platform

4. MT will become just one part of a much larger AI offering

5. Translation data and metadata will be considered a critical asset, giving today’s tech-savvy language professionals a very bright future
3. Multimodal Foundation Models

The next platform for MT ... and so much more

- A paradigm shift in AI
- Massive pre-trained models provide the foundation for task-specific models
- Initial training is unsupervised and can include all kinds of data but does not include a specific task
- Can be “trained” for a downstream task by simply providing it with a natural language description of the task
- Spooky: this ability to do tasks based on verbal prompts was neither specifically trained for nor anticipated to arise.
ChatGPT

- Developed by OpenAI
- Released in November 2022
- Based on GPT-3
  - Transformer Neural Network
  - “Generative Pre-trained Transformer”
  - Trained with massive
  - LLM with 175 billion “parameters”
- Generates text in response to prompts

PS: GPT-4 is being released next week.

Write a sonnet about post-editing machine translation

Amidst the whirring gears and hum of steel,
A marvel of technology resides,
A tool to make our language barriers yield,
And bring our tongues and cultures side by side.

Yet though its speed and reach we all admire,
And welcome the relief it brings us all,
There still remains a need for human fire,
To give the text that certain human drawl.

For only in the mind of man, we find,
The depths of nuance, humor, and emotion,
A richness that machines can’t quite unbind,
A certain je ne sais quoi of expression.

So let us hail this wondrous, modern aid,
And wield it deftly with our human blade.
How will foundation models be applied to projects?

Part 4: Simplified view of a workflow with AI-assisted content creation, MT, QE, and post-editing
How will foundation models be applied to projects?

Part 5: Simplified view of a workflow *with all kinds of Al-driven functionality*
A 100-year-old lesson about technology-driven change

- At the first Academy Awards (1929), all of the nominees for Best Picture were silent movies.
At the first Academy Awards (1929), all of the nominees for Best Picture were silent movies.

At the second Academy Awards (1930) none of them were.

Actors unable to transition to the world of talkies lost work, including Charlie Chaplin, Rudolf Valentino, Mary Pickford, Lillian Gish, Clara Bow.

But many, many new jobs were created.
What kind of jobs will be created by AI in our industry?

- Data Curator
- Data Scientist
- Terminologist
- Corpus Linguist
- Computational Linguist
- Language Engineer
- Communication Analyst
- Semantic Analyst
- Prompt Designer
- Translation Technology Expert
- Language Technology Analyst
- Language Process Analyst
- Machine Learning Evaluator
- AI Ethicist
- Translation Quality Assessor
- “Gatekeeper” Translator
- “Gatekeeper” Interpreter
Data Curator

Tasks:

• Ensure data quality and consistency through the implementation of data governance practices

• Define and enforce data styling and terminology standards

• Develop and maintain a data catalog to ensure efficient data discovery and accessibility

• Work with stakeholders to understand data requirements and ensure data availability and accuracy

• Collaborate with data scientists and machine learning teams to support the development of new models and features

• Manage data retention and archival policies to ensure compliance with data privacy regulations

Requirements:

• Experience in data management and data governance

• Familiarity with machine learning and data science methodologies

• Strong understanding of data storage and retrieval systems

• Experience with SQL and NoSQL databases

• Excellent written and verbal communication skills

• Ability to work with cross-functional teams and manage stakeholder relationships

• Strong problem-solving and analytical skills

• Bachelor's or Master's degree in Computer Science, Data Science, or a related field
Machine Learning Evaluator

Tasks:

- A Machine Learning Evaluator evaluates the performance and effectiveness of machine learning models, particularly Neural Machine Translation (NMT) and Large Language Models.
- Develop and implement testing methodologies to evaluate the performance and accuracy of NMT and Large Language Models.
- Analyze and interpret the output of machine learning models to identify strengths, weaknesses, and areas for improvement.
- Collaborate with engineering and research teams to provide recommendations and feedback for improving machine learning models.
- Conduct experiments and A/B testing to evaluate the impact of changes to machine learning models.

Requirements:

- Bachelor's or Master's degree in Computer Science, Mathematics, or a related field.
- 2+ years of experience in machine learning evaluation or related field.
- Strong understanding of machine learning concepts and techniques.
- Experience with NMT and Large Language Models.
- Proficiency in programming languages such as Python or R.
- Familiarity with machine learning tools and frameworks such as TensorFlow, PyTorch, and Keras.
- Strong analytical and problem-solving skills.
- Excellent communication and collaboration skills.
Gatekeeper Translator

Tasks:

• Ensure that automatically translated texts meet quality standards for their respective use cases

• Sample translations using traditional language quality assessment techniques

• Ensure compliance with corporate style and terminological guidelines

• Evaluate the suitability of translations for the intended audience

• Provide feedback to the translation team to improve the quality of future translations

• Stay up-to-date with the latest advances in translation technology and quality assessment techniques

Requirements:

• Bachelor’s or Master’s degree (or equivalent experience) in Translation, Modern Languages, Linguistics, or a related field

• Strong background in translation and language quality assessment

• Excellent written and verbal communication skills in source and target languages

• Ability to work with cross-functional teams and manage stakeholder relationships

• Strong problem-solving and analytical skills

• Familiarity with machine learning and NLP techniques
Language Engineer

Tasks:
• Design and implement AI models to process and understand multimodal language data, including translation memories, monolingual texts, voice recordings, and video transcriptions.
• Develop and implement algorithms for natural language understanding, natural language processing, and natural language generation.
• Work with cross-functional teams to understand customer requirements and design solutions that meet those needs.
• Collaborate with data scientists and engineers to ensure high quality and accuracy of language models.
• Keep up to date with the latest developments in AI and natural language processing.

Requirements:
• Strong programming skills in Python and/or other programming languages.
• Experience with NLP libraries such as NLTK, spaCy, or PyTorch.
• Strong understanding of natural language processing, including syntax, semantics, and pragmatics.
• Experience working with machine learning algorithms, including deep learning, decision trees, and random forests.
• Excellent written and verbal communication skills.
• Bachelor's or Master's degree in Computer Science, Linguistics, or a related field.
We’ve already come so far

An IBM Electronic Calculator speeds through thousands of intricate computations so quickly that on many complex problems it’s like have 150 EXTRA Engineers. No longer must valuable engineering personnel ... now in critical shortage ... spend priceless creative time at routine repetitive figuring. Thousands of IBM Electronic Business Machines ... vital to our nation's defense ... are at work for science, industry, and the armed forces, in laboratories, factories, and offices, helping to meet urgent demands for greater production.

1953 advertisement for the IBM 701 mainframe
Is there an enemy here?

You are not in competition with AI.

You are in competition with people who use AI better than you do.
What strengths will be required to thrive in these jobs?

• All the great skills you already have
  • Subtle and sophisticated knowledge of language
  • Excellent proficiency in two or more languages
  • No fear of other languages
• Deep curiosity
• An appreciation for “algorithmic thinking”
• Comfort level (or better) with data and databases
• The audacity to work on skills that will make the old you redundant
Sure, but what else?

Imagination
Thank you for your attention!
Let’s keep the conversation going!

https://www.linkedin.com/in/jaymarciano/