CDISC SDTM CT 翻译更新进展

Lenny Li 李钰琳 2024/11/08

Lenny Li

李钰琳

工作经验：13年以上CRO和药企工作经验，负责多个治疗领域和适应症的产品的I - IV期临床试验统计编程，熟悉监管机构临床数据递交规范，熟悉CDISC标准，SAS编程熟练

爱好：瑜伽，旅行，咖啡，音乐，阅读

个人声明

以下仅为个人观点，不代表公司立场。

CDISC SDTM CT 翻译既往概况

- Since 2016年
  - 初稿 (未发布)

- 2021年7月
  - 发布《Most Commonly Used Controlled Terminology - Chinese Translation 常用受控术语》
  - 基于SDTM Terminology 2021-03-26
  - 418个中文术语
  - 24个代码表

https://www.cdisc.org/translations/chinese

CDISC SDTM CT 翻译更新

- 2023年
  - 基于SDTM Terminology 2022-09-30
  - 共38953个术语
  - 16033个术语已翻译/无需翻译 (如--TESTCD)
  - 13694个非常用术语暂不翻译
  - 9226个术语纳入2023更新
    - 59个代码表
  - 未定稿

CDISC SDTM CT 翻译更新 (Cont.)

- 2024年
  - 基于SDTM Terminology 2024-09-27
  - 共43976个术语
  - 6282个新术语 (vs SDTM Terminology 2022-09-30)
    - 158个新代码表, 129个原有代码表
    - 868个新术语纳入2024更新
      - 48个原有代码表
  - 进行中

CDISC SDTM CT翻译组成员 (排名不分先后)

- Lead：Lenny Li 李钰琳
- Co-lead：Sharon He 何思韵
- 志愿者：Aurora Yu, 常杰, 陈汝洁, 方焱, 韩琨, 棘玉荣, 姜鑫, 李璠宇, 李晓帆, 刘慧,刘睿颖, 骆豪,牛喜草, 潘云蕾, 秦莉莎, 苏国娜, 孙延春, Vera Zhu, 王鹏飞, 肖越, 杨济豪, 杨凌, 张柳, 张卢, 张秀萍, 赵翔, 朱潇琳
- 顾问：Victor Wu 吴崇胜, John Wang 王军

Thanks!

翻译更新流程

1	2	3	4	5
AI翻译 	志愿者独立review 	翻译更正 	一致性检查 	Final or 待定 

困难与挑战

- 不同代码表涉及的领域跨度大
  - 如实验室检查项, 解剖学位置, 药物剂型等

- 部分关键词、医学术语无中文翻译, 或无文献来源
  - 如实验室检查项:
    - Promyeloblasts (? ?) are the earliest stage of granulocyte development. They are large cells with a high nuclear-cytoplasmic ratio and prominent nucleoli*.
    - Myeloblasts (原始粒细胞?) are slightly more mature than promyeloblasts. They are still large cells but have less prominent nucleoli and begin to show some granulation*.
    - Promyelocytes (早幼粒细胞?) are the next stage after myeloblasts. They are characterized by the presence of primary (azurophilic) granules in the cytoplasm*.
    - Myelocytes (中幼粒细胞?) are more mature than promyelocytes. They have secondary (specific) granules and a lower nuclear-cytoplasmic ratio*.
    - Metamyelocytes (晚幼粒细胞?) are the stage after myelocytes. They have a kidney-shaped nucleus and are more mature, resembling mature neutrophils*.

**From Copilot, not verified*

困难与挑战 (Cont.)

- 部分长术语无中文翻译
  - 如解剖学位置:
    - CIRCUMFLEX, OBTUSE MARGINALS, LEFT POSTEROLATERAL AND LEFT POSTERIOR DESCENDING ARTERY BRANCHES (回旋支, 钝缘支, 左后外侧和左后降支动脉?)
    - MID/DISTAL LEFT ANTERIOR DESCENDING CORONARY ARTERY AND ALL DIAGONAL CORONARY BRANCHES (中段/远段左前降支冠状动脉和全部对角支动脉?)
    - RIGHT CORONARY ARTERY, RIGHT POSTERIOR DESCENDING, RIGHT POSTERIOLATERAL AND ACUTE MARGINAL BRANCHES (右冠状动脉, 右后降支, 右后外侧和锐缘支动脉?)

- 字面翻译 vs 语义翻译
  - 如相关发现Test名称:
    - Occurrence Indicator: 发生指示符, 发生指标, **是否发生?**
  - 如肿瘤疗效评估Test名称:
    - Best Overall Response: 最佳总体应答, 最佳总体反应, **最佳总体疗效?**
    - Bone Marrow Response: 骨髓应答, 骨髓反应, 骨髓缓解, **骨髓疗效?**
  - 如试验概要参数Test名称:
    - Trial Length: 试验长度, **试验时长?**

困难与挑战 (Cont.)

- 同一术语在不同代码表中翻译不一致
  - 如肿瘤疗效评估结果 vs 肿瘤或病灶特性Test结果:
    - INCREASED: 增加, 增多, 增大, 扩大, 加重
    - DECREASED: 减少, 减小, 缩小, 减轻

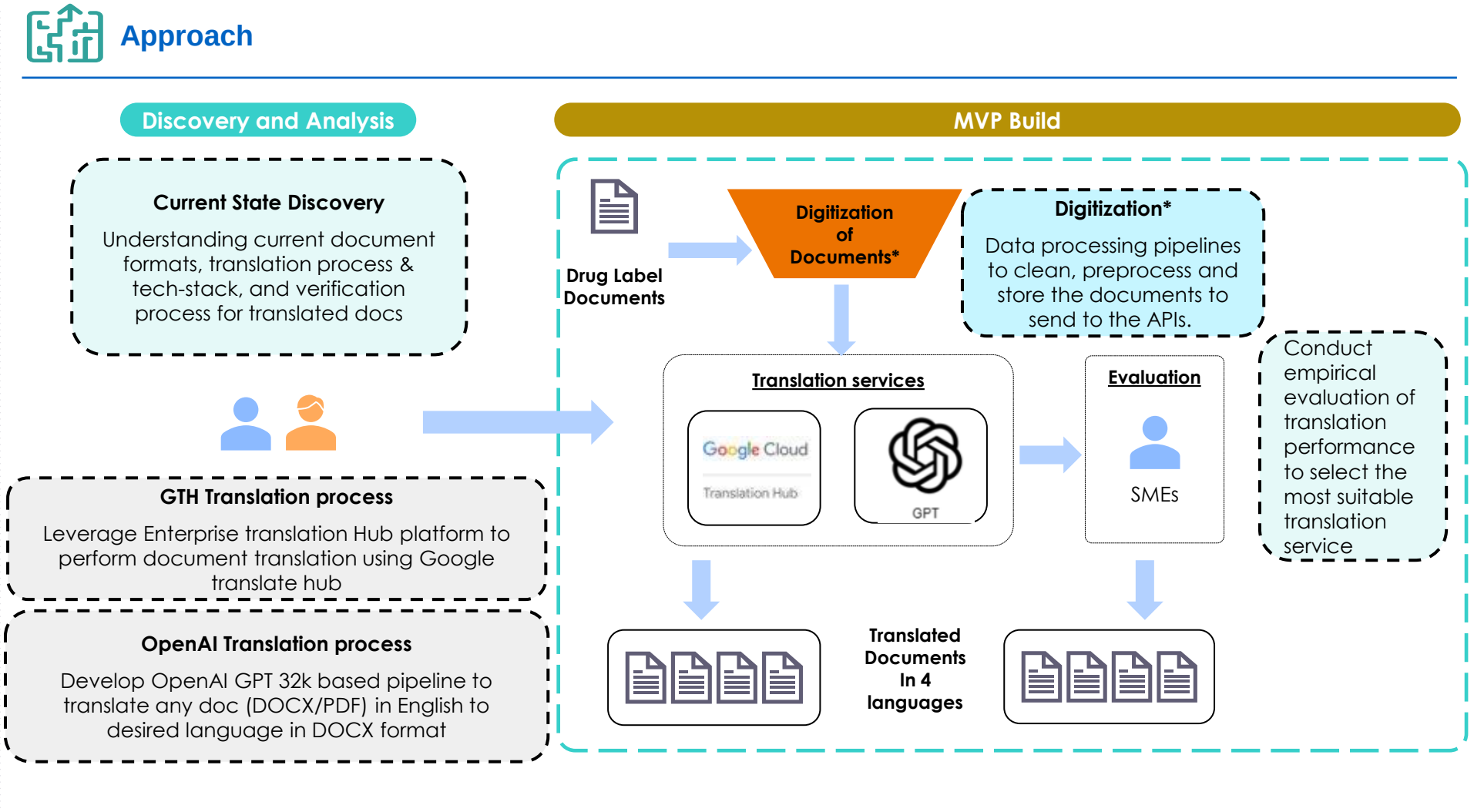
- 同一关键词在不同术语中翻译不一致
  - 如实验室检查项:
    - Total Cells: **总细胞**，总细胞计数，细胞总数
    - Band Form: 带状核, 杆状核, 带型

解决方案

- 继续探索AI翻译工具
- 搜索中、英文医学词典和文献
- 长术语拆分关键词，翻译关键词后连成句
- 缩写、人名、专有名词等不翻译
- 保持翻译一致性
- 产学研合作

Exploring the Use of LLMs for Language Translation

Approach



Exploring the Use of LLMs for Language Translation (Cont.)

Highlights

Google translation service emerges as the suitable candidate for translation based on the empirical evaluation

- Successfully demonstrated (live-demo) model performance and compared outputs against golden dataset
- Improved performance and outputs of the LLMs through multiple approaches and strategies

29
Translation outputs

6
Language dialects

~8
SME feedback sessions

>7-day improvement
In time to generate initial draft of the translation

Project Deliverables

- Created codebase using Open AI GPT-4 32k to translate DOCX/PDF documents from English to 6 language dialects with:
  - PDF / DOCX compatibility
  - Glossary support

- Machine evaluation codebase was also developed to provide automated GEMBA & embedding based similarity scores for the translated documents

- Collected SME feedback on the translations from Open AI and Google translation service

Google Translation vs OpenAI GPT-4

Sl. No.	Language	Approach	Average GEMBA (out of 100)	Average Embedding based similarity (out of 1.0)	Average Fidelity rating (a scale of 1 to 7, 7 being best)	Average Fluency rating (a scale of 1 to 7, 7 being best)	Average Format rating (a scale of 1 to 7, 7 being best)	Overall quality for majority of docs. (High / Medium / Low)
1	Simplified Chinese	Google translation service	95	0.4	4	4	5	Medium
2		OpenAI GPT-4 32k model	89	0.3	3	3	2	Low

Sl. No.	Language	Approach	Overall Feedback	Example Issues
1	Simplified Chinese	Google translation service	• **Translation struggled with medical terms** • Otherwise, happy with Google translations; believes a glossary for medical terms would correct major remaining issues	• Incorrectly translated certain phrases
2		OpenAI GPT-4 32k model	• **Formatting** made documents difficult to assess; in particular, made the document non-compliant with Chinese regulations • Compared to Google, **struggled more with medical terms**	• Duplicated certain phrases • Misplaced some translations on page • Some translations were difficult to understand

Q & A

Thank You