SEMALYTIX

KARLSRUHE, SEPTEMBER 9TH 2019

SEMALYTIX Prêt-à-LLOD

Multilingual Text Analytics for Extracting Pharma Real-World Evidence

Semalytix Pilot within
EU Horizon 2020 Project Prêt-à-LLOD

Matthias Orlikowski, Susana Veríssimo, Matthias Hartung

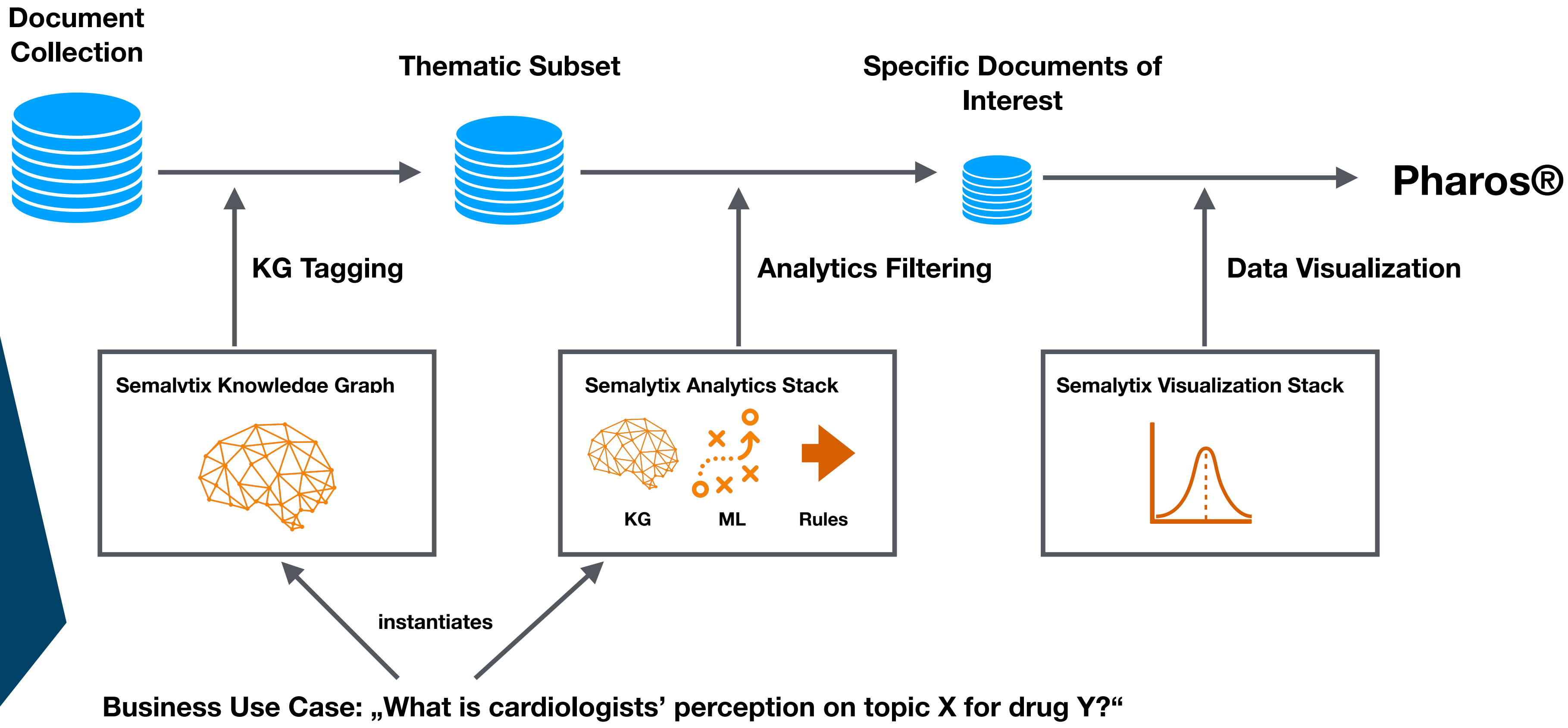
Twitter @morlikow, @semalytix

Outline

1. **Background and Motivation**
2. Case Study I: Machine Translation Pipeline (English-Japanese)
3. Case Study II: Cross-Lingual Transfer (English-Spanish)
4. Conclusion

Tech Overview

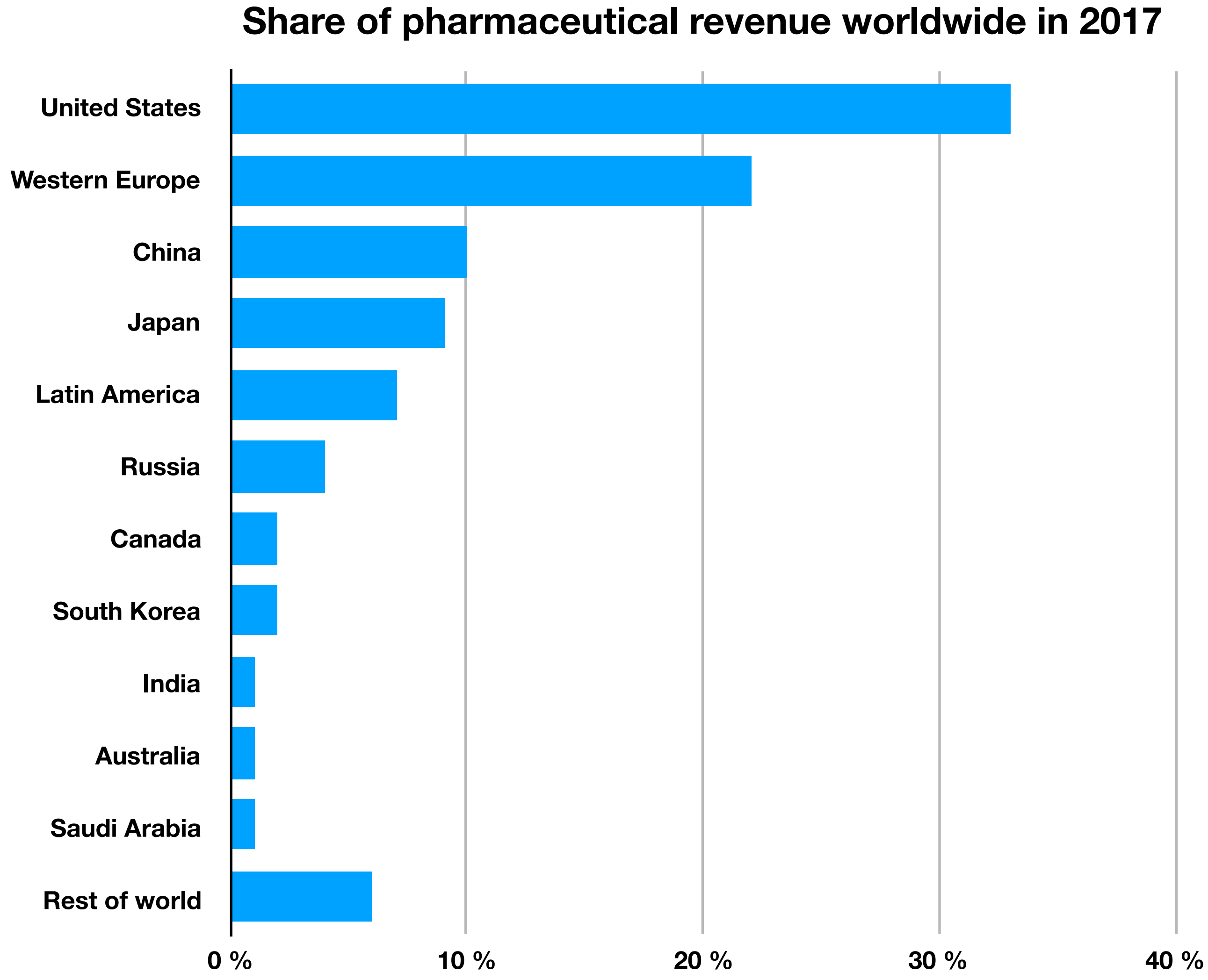
Semalytix applies a combination of Semantic Technologies, Machine Learning and Natural Language Processing in order to automatically understand pharma-specific text at scale



Document Collection
Thematic Subset
Specific Documents of Interest
Pharos®
KG Tagging
Analytics Filtering
Data Visualization
Semalytix Knowledge Graph
Semalytix Analytics Stack
KG
ML
Rules
Semalytix Visualization Stack
instantiates
Business Use Case: „What is cardiologists' perception on topic X for drug Y?"

SEMALYTIX

Motivation - Pharma is global



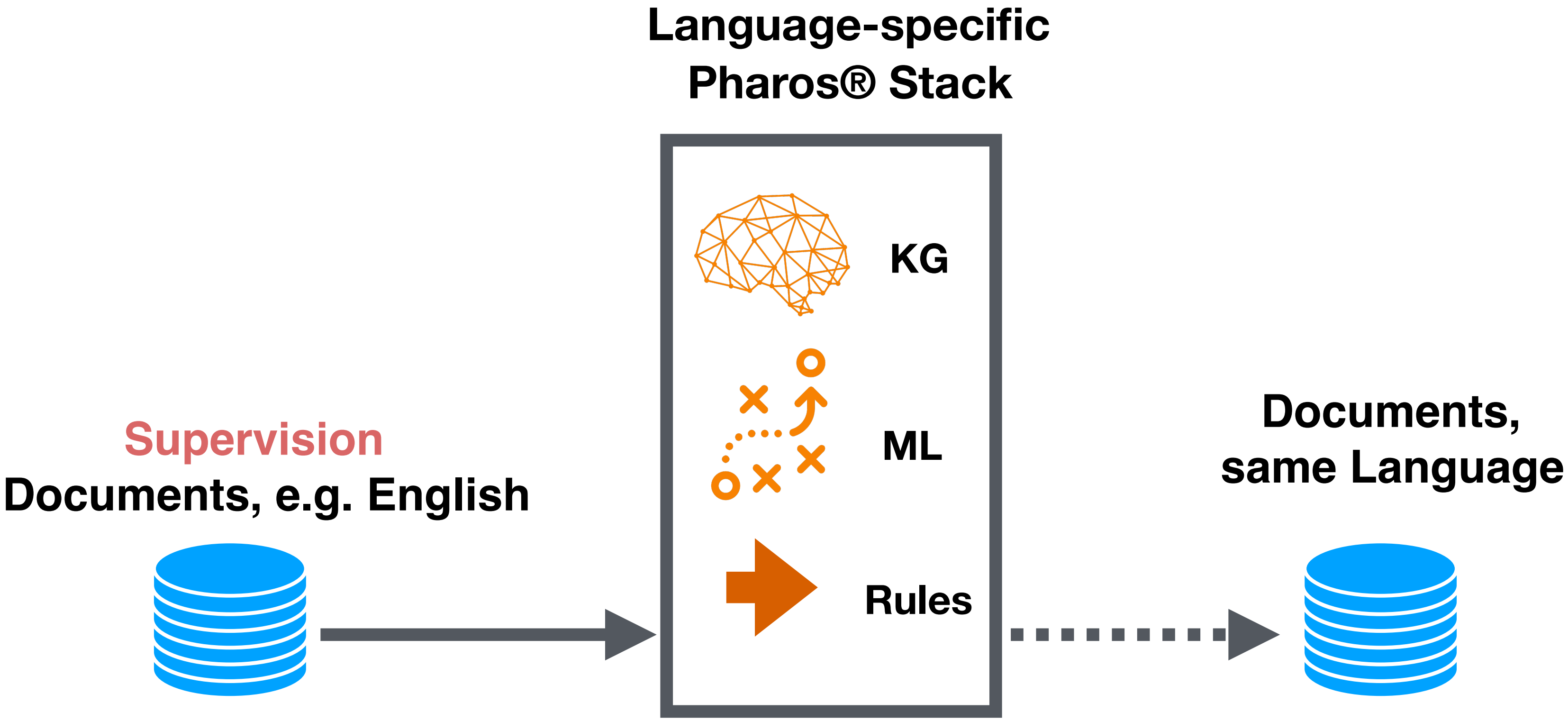
source:

Our customers are pharmaceutical companies operating in **global markets**.

A large share of pharmaceutical revenue is generated in **non-English markets**.

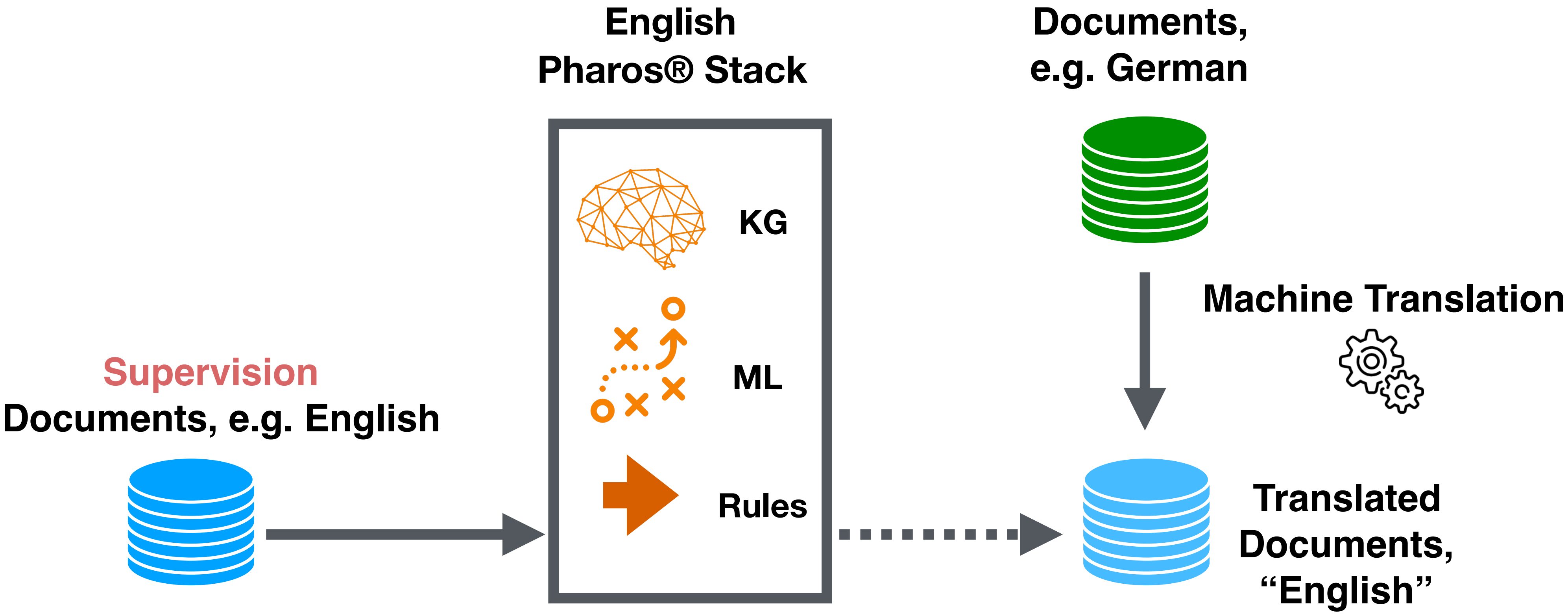
Answering business questions in a meaningful and actionable way requires **multilingual text analytics**.

Multilingual Analytics - Building from Scratch



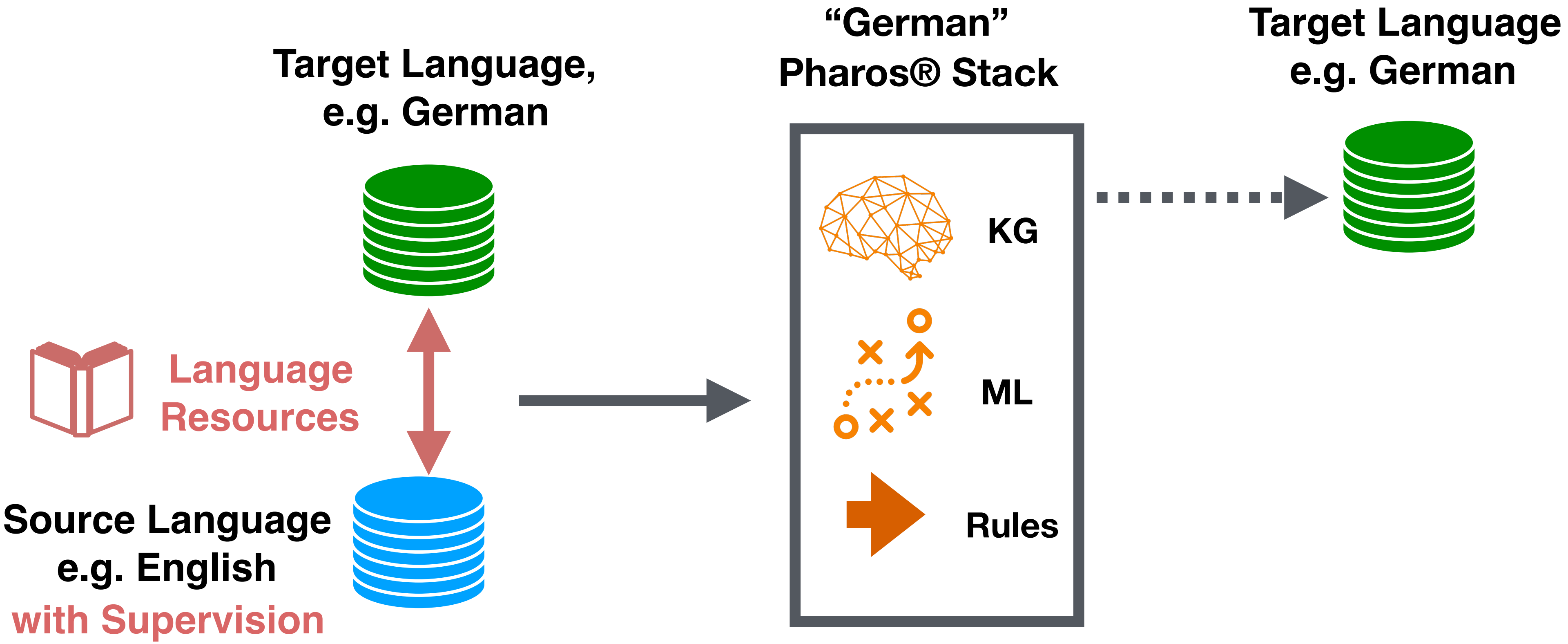
- Building a stack of analytical components needs some degree of **domain knowledge / supervision**
- Result is language-specific: More languages? Rebuild the stack for each!

Multilingual Analytics - Machine Translation Pipeline



- **Low-effort approach** to analyse non-English text without the need for supervision in language of interest
- Potential performance gaps resulting from surface translation pipeline could be mitigated by **language-specific model optimisations**

Multilingual Analytics - Cross-lingual Transfer



- Build stack **for the target language** via cross-lingual transfer, with the additional opportunity of task- and domain-specific optimization
- Comparatively low-effort as **no supervision for target language** required

Prêt-à-LLOD

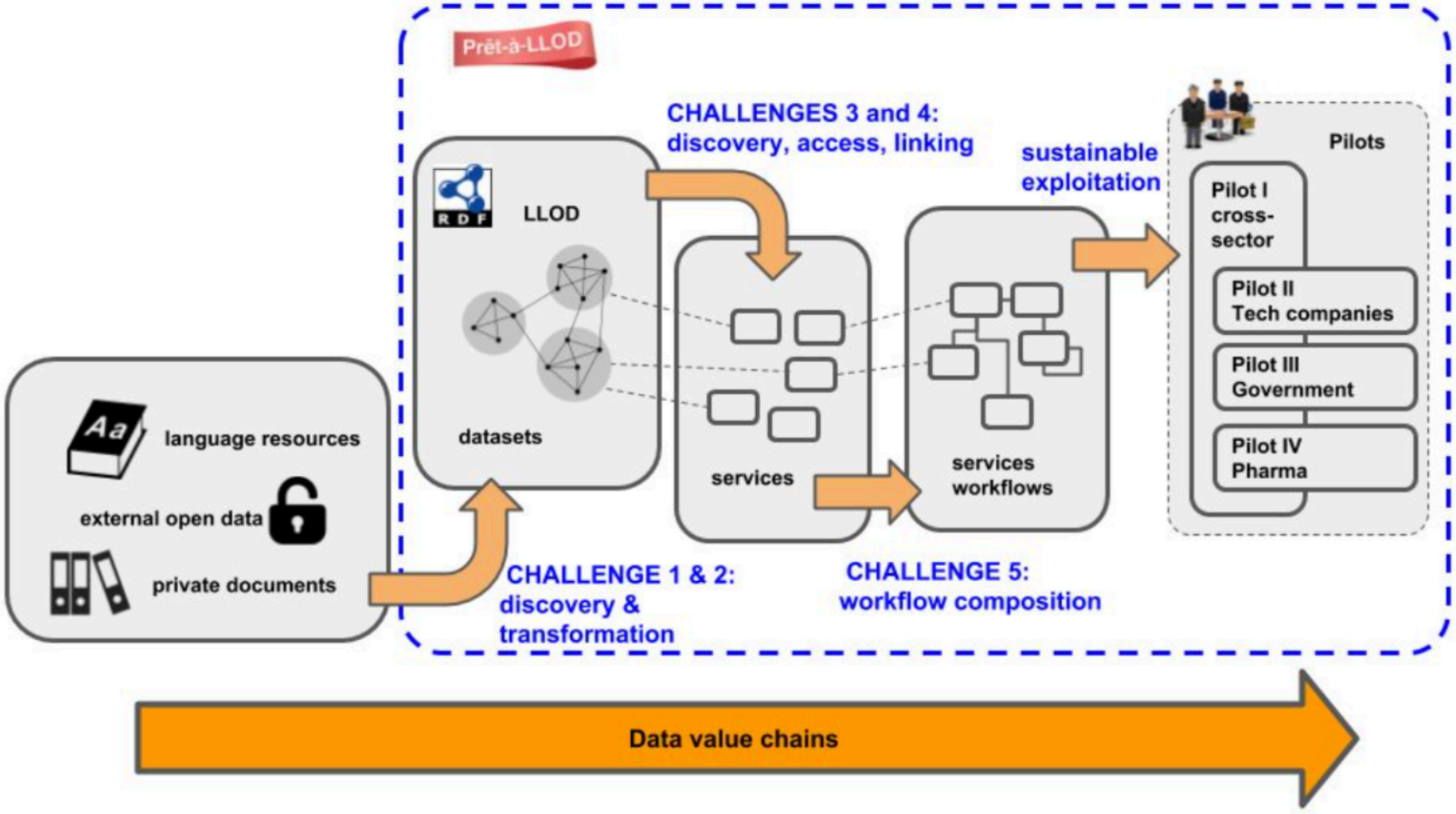


Figure 1.4: Different research challenges along the data value chains in Prêt-à-LLOD

Outline

1. Background and Motivation
2. **Case Study I: Machine Translation Pipeline (English-Japanese)**
3. Case Study II: Cross-Lingual Transfer (English-Spanish)
4. Conclusion

Case Study I - Machine Translation Quality

Machine Translation Services considered for Language Pair JP-EN:

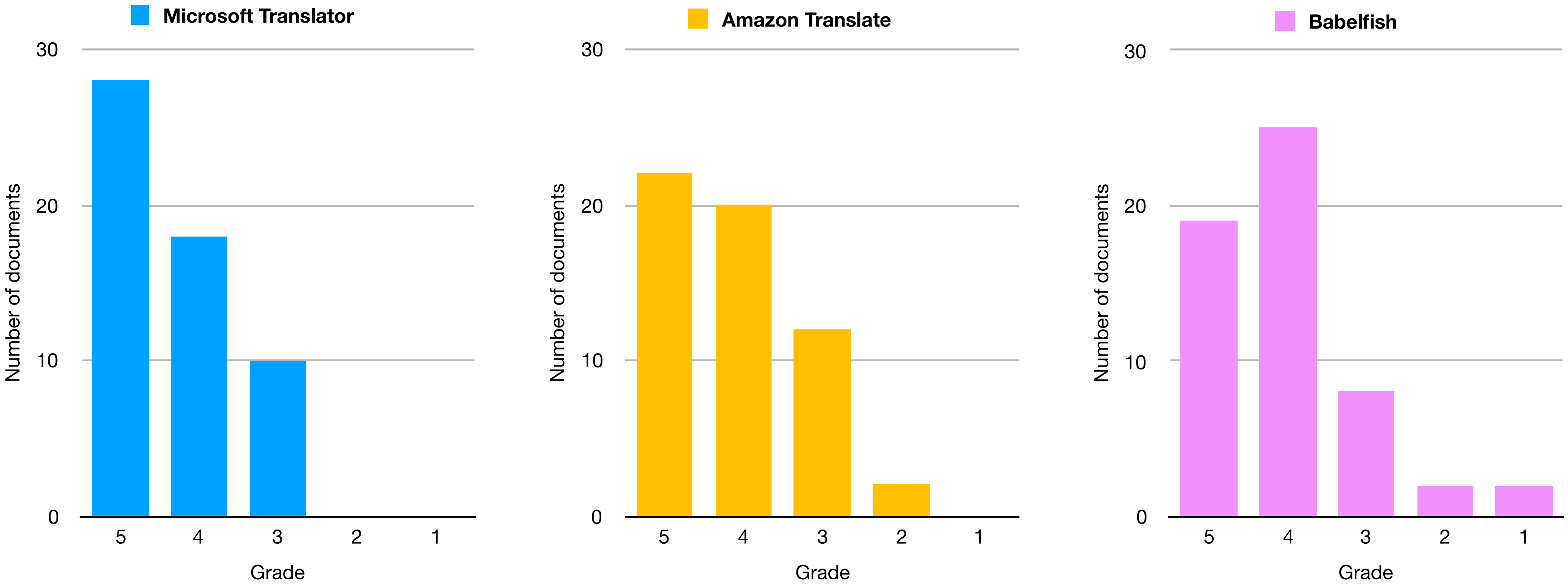
- **Amazon Translate**, https://aws.amazon.com/translate/
- **BabelFish**, https://www.babelfish.com/
- **Google Translate**, https://translate.google.com
- **Microsoft Translator**, https://www.bing.com/translator
- **Reverso**, http://www.reverso.net/text_translation.aspx
- **Systran**, https://translate.systran.net/translationTools/text

Evaluation Procedure:

- Small sample of sales interaction documents was translated using each machine translation service
- Each translated text was rated based on its coherence from 1 („completely incomprehensible") to 5 (“good English”)
- Main goal: inform selection of a translation service for evaluating Pharos® analytical components on documents translated from Japanese
- We plan to carry out more in-depth evaluations of translation quality with native speakers and/or professional translators

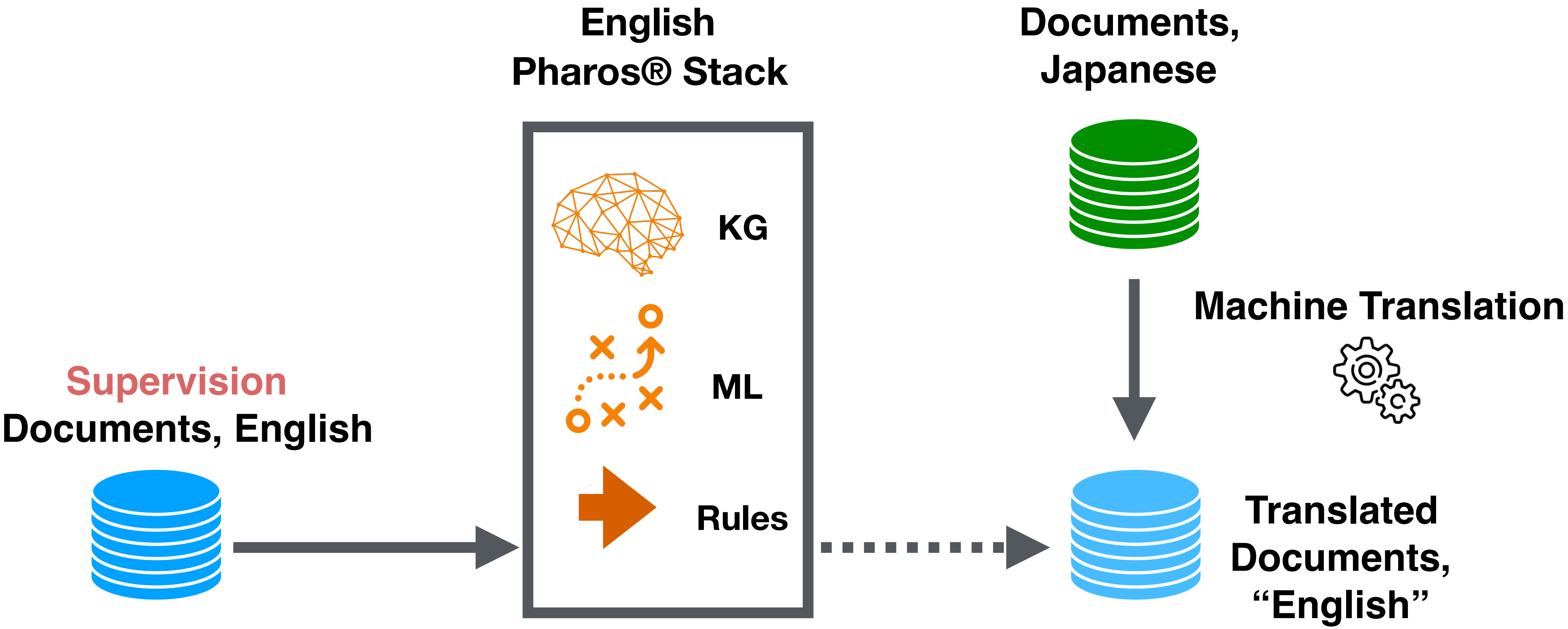
Case Study I - Machine Translation Quality

Results for Top 3 Translation Services:



Microsoft Translator produces the **best quality translations** for Japanese documents in our genre, with a large number of high-quality translations and no poor-quality translations.

Case Study I - Machine Translation Pipeline (EN-JP)

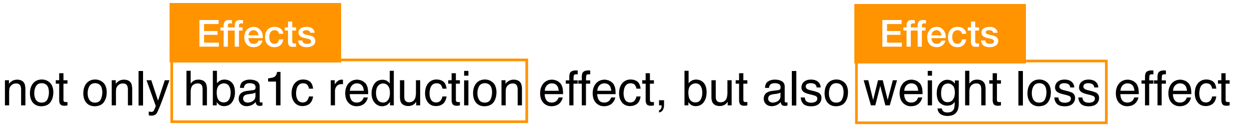


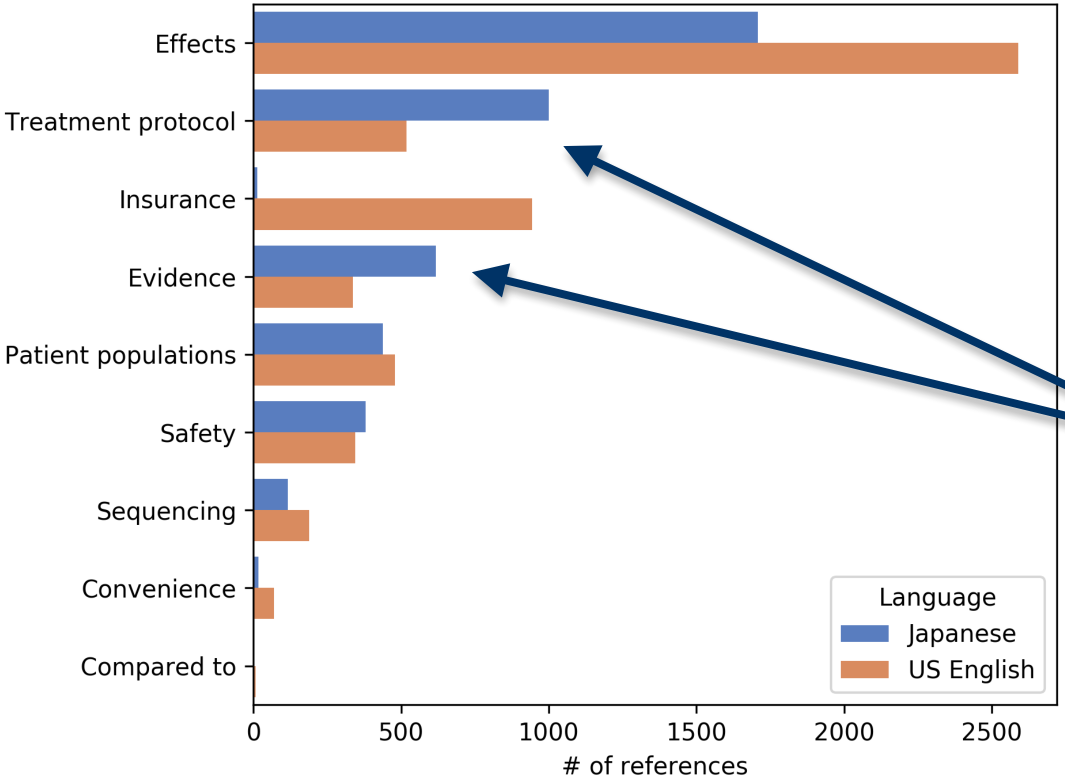
- A same-sized sample from sales interaction documents for both **Japanese and English**
- Japanese texts translated into English using the **Microsoft Translator API**
- Selection of example analytical components for comparison:
  - **Sentiment classification, Topic detection, Entity tagging**

Case Study I - Topic Detection

Topic detection extracts parts of texts which refer to one of a number of salient concepts that are recurrent across many documents in the corpus.

HbA1c低下作用だけでなく、体重減少効果もある → not only hba1c reduction effect, but also weight loss effect

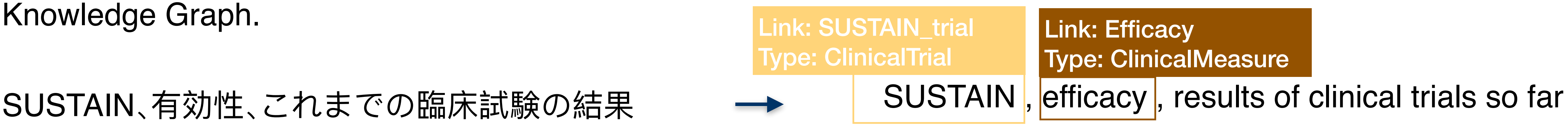




- Majority of topics with high salience in EN also **robustly detected** in JP texts.
- As expected: Similar or **higher coverage** of concepts in **EN data**, as the detection was tuned for this language.
- For "**Treatment protocol**" and "**Evidence**", more references can be found in Japanese, suggesting **particular richness** for these concepts.
- Some concepts are very market-specific by their nature (e.g.,"Insurance"), thus leading to few finds in the translations.

Case Study I - Entity Tagging

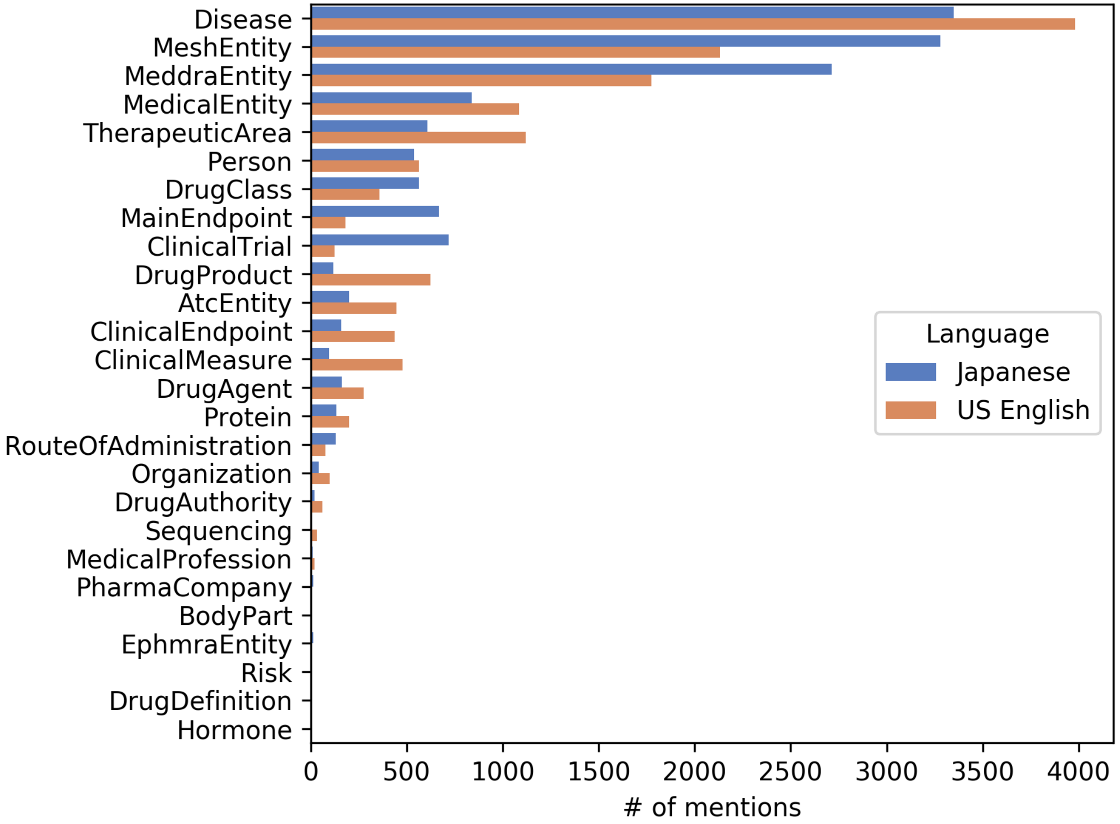
Entity tagging detects mentions of biochemical, pharmaceutical and medical entities like drug products, clinical trials and diseases in text. Mentions are linked to entity concepts in the domain-specific Semalytix Knowledge Graph.


SUSTAIN、有効性、これまでの臨床試験の結果

Link: SUSTAIN_trial
Type: ClinicalTrial

Link: Efficacy
Type: ClinicalMeasure

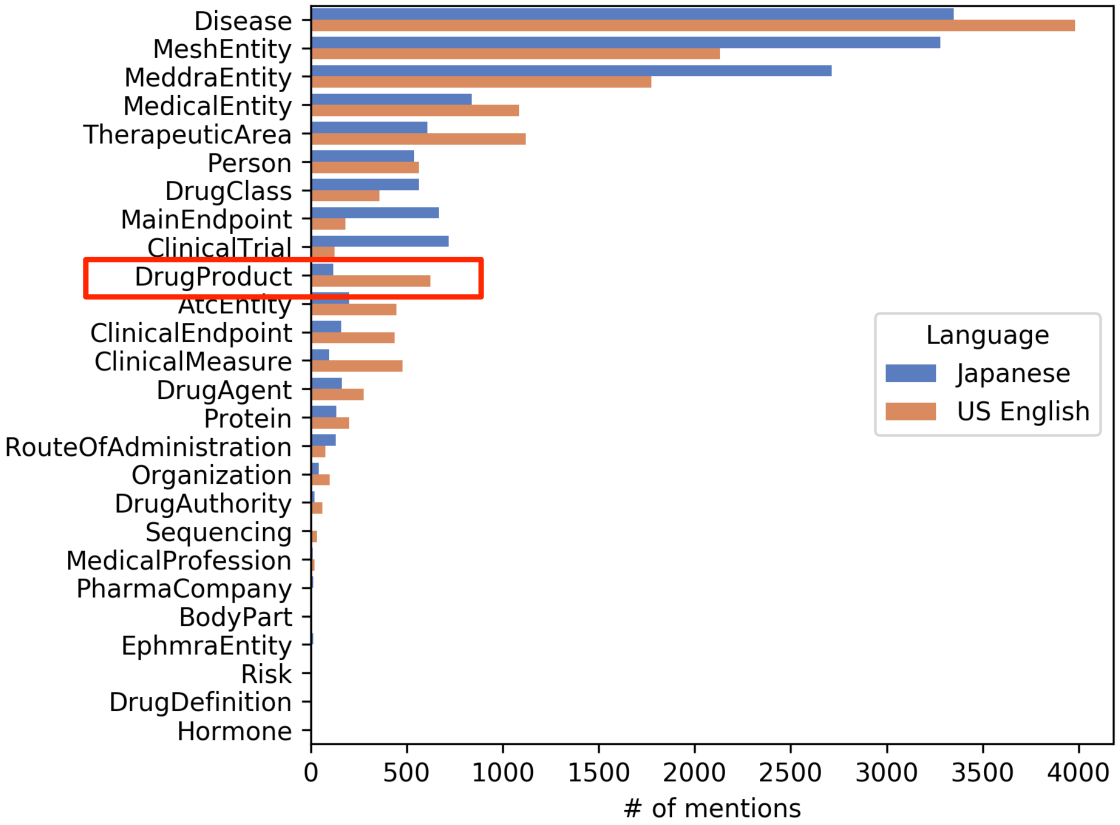
SUSTAIN, efficacy, results of clinical trials so far



- Majority of entities recognized in EN also **robustly detected** in JP texts.
- As expected: higher coverage in EN texts for most entity types, as the tagging was tuned for this language.
- However, for entity types from the MeSH and MedDRA ontologies, we even find more references in texts translated from JP

Case Study I - Entity Tagging, Negative Example

Negative example: What is the reason for the low coverage of DrugProducts in JP?
As we selected interactions about specific drug products and tagging performance is highly robust in EN, we would expect to detect DrugProducts mentioned frequently in JP texts as well.



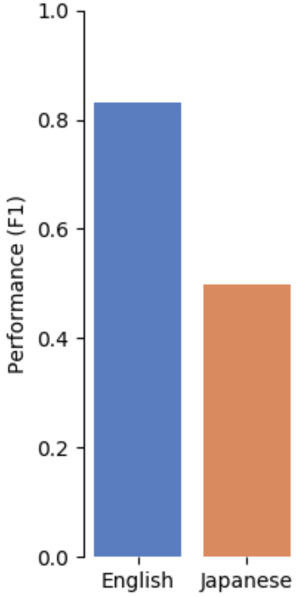
- Japanese frequently uses Latin script for foreign **technical terms**, but we find a number of drug products mostly written in a syllabic script.
- An error analysis reveals many **erroneous transliterations** of drug products introduced by the domain-agnostic translation service.
- Similar problems may occur for **all entity types**; the issue can be mitigated by tuning entity taggers for distant spelling variations, based on **language-specific knowledge graph enrichment**.

Case Study I - Sentiment Analysis

Pharma-specific sentiment analysis detects the sentiment of a text, taking into account the positivity and negativity associated with terms and phrases in the pharma/medical domain and specific contexts.

心血管ｲﾍﾞﾝﾄ減少が証明された。 → Cardiovascular events have been proven to decrease

Positive



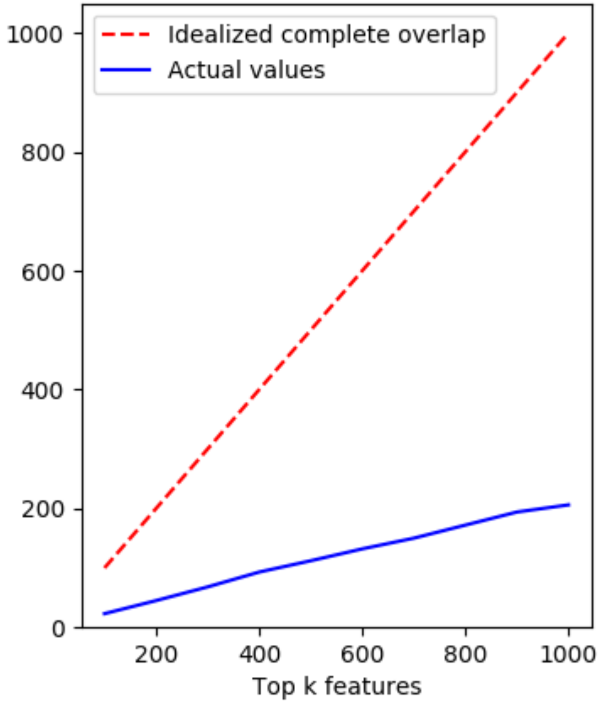
Results for applying sentiment analysis (tuned on English texts) on translated documents show a large **performance gap (-40%)** in comparison to performance on English original text.

What explains the observed performance gap?

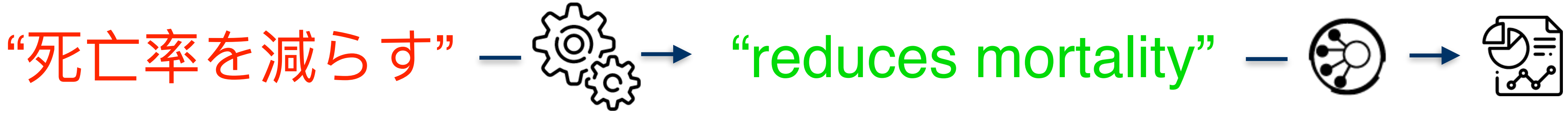
Feature selection: We extract the top *k* features showing strongest association with positivity in EN data.

Comparison: Only a small proportion of these important features (approx. 20%, across different values of k) occurs within the sample of positive documents translated from Japanese.

Conclusion: The translated data exhibits largely different markers of positivity from what the model has learnt on original English data.



Case Study I - Results Overview

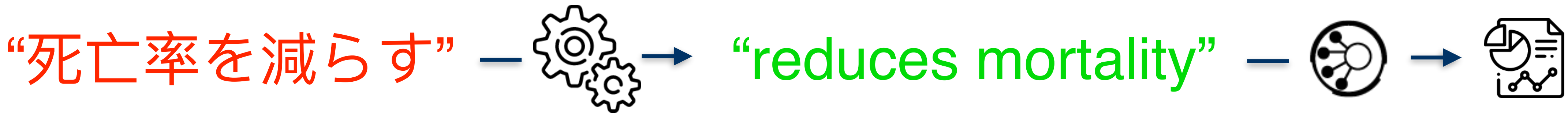


“死亡率を減らす” → “reduces mortality” →

Results Overview:

- Surface translation can be considered sufficiently robust to provide an **informative overview** about topics in Japanese texts
- Automated surface translation enables Pharos® analytics capacities at **high precision** for topic detection and entity tagging in Japanese.
- With respect to **coverage**, as well as accuracy of sentiment models, performance lags behind Pharos® standards established for English

Case Study I - Limitations



Japanese and English are a rather distant language pair:

- Different **script**: Japanese uses kanji (Chinese characters), two syllabic scripts (may cause issues of transcription and transliteration)
- Different **word order**: subject–object–verb (SOV), equivalent to “Mary apples likes”
- Different **grammatical properties**: nouns have no grammatical number and gender, no articles
- Differences in **cultural context**

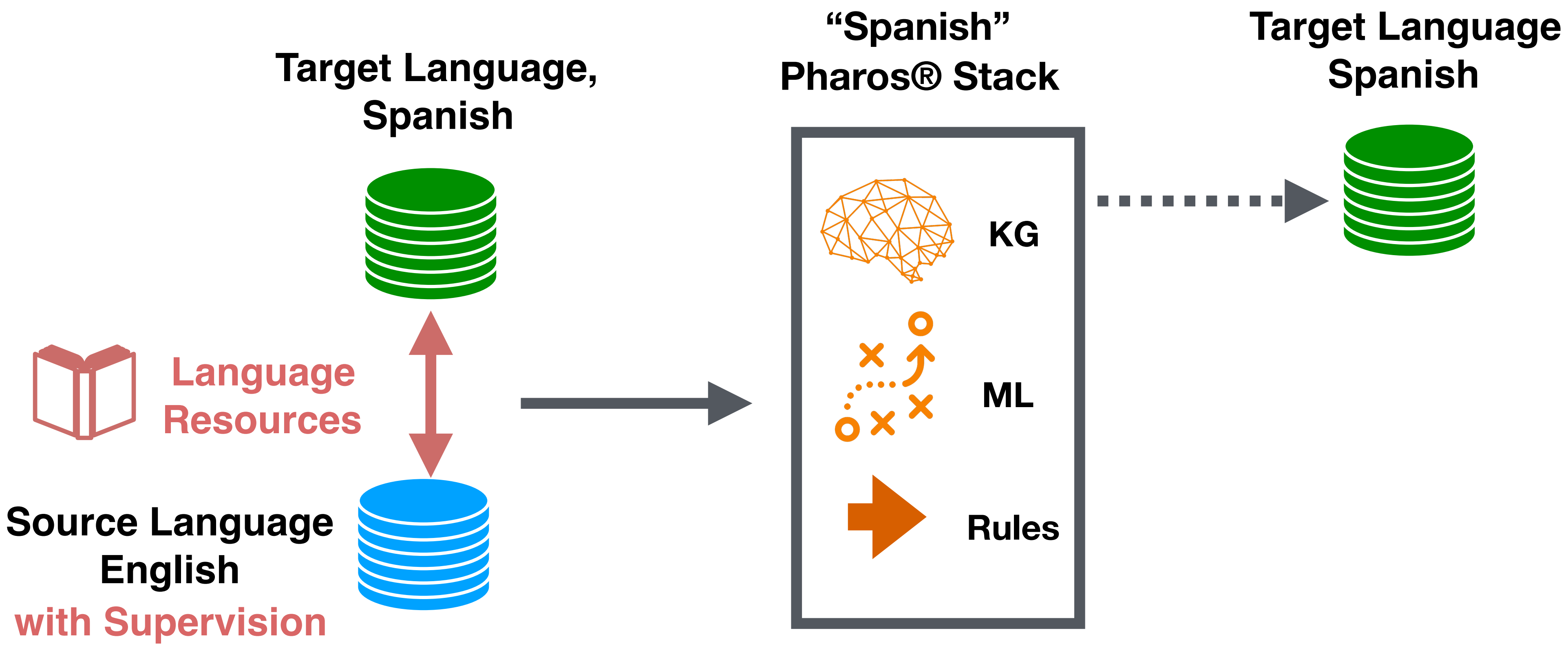
Surface translation imposes inherent limitations on down-stream analytics:

- Domain specifics of pharma-related text (technical terms often not sufficiently covered in translations)
- Heavy use of fragmented, abbreviated language in our type of text
- Automatic translations may contain deviations from standard English
- Existing Pharos® analytical components have not been tailored to deal with translated text - but is that what we want to do?

Outline

1. Background and Motivation
2. Case Study I: Machine Translation Pipeline (English-Japanese)
3. **Case Study II: Cross-Lingual Transfer (English-Spanish)**
4. Conclusion

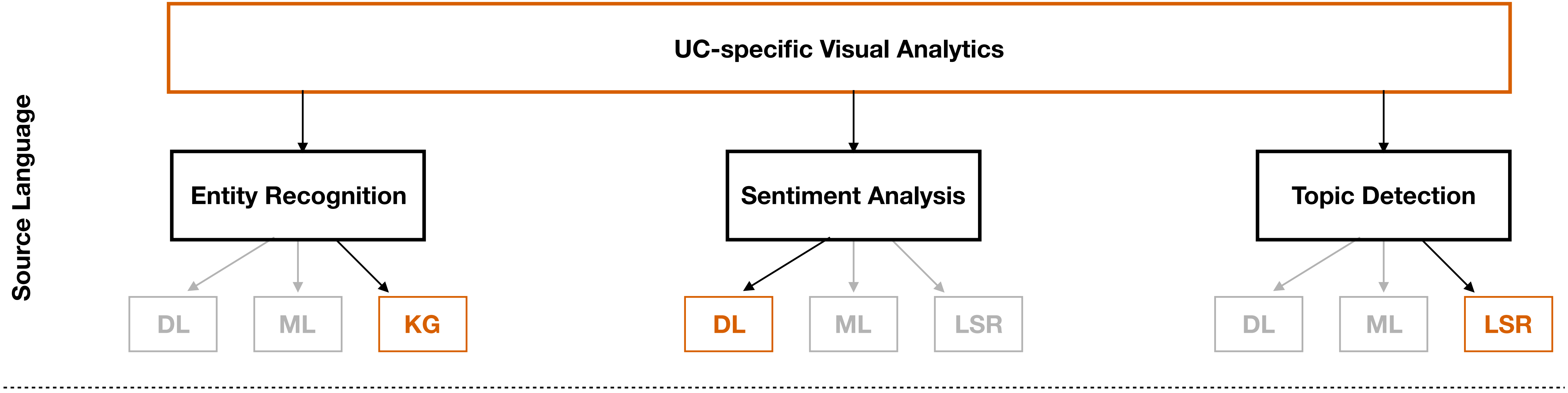
Case Study II - Cross-Lingual Transfer (EN-ES)



- Small sample (n=1000) from sales interaction documents for both **Spanish and English**
- Prototype based on *Bi-Lingual Sentiment Embeddings: Joint Projection of Sentiment across Languages* (Barnes et al., 2018) (**BLSE**), part of Thesis project by Susana Veríssimo
- Evaluate impact on performance of **different language resources**

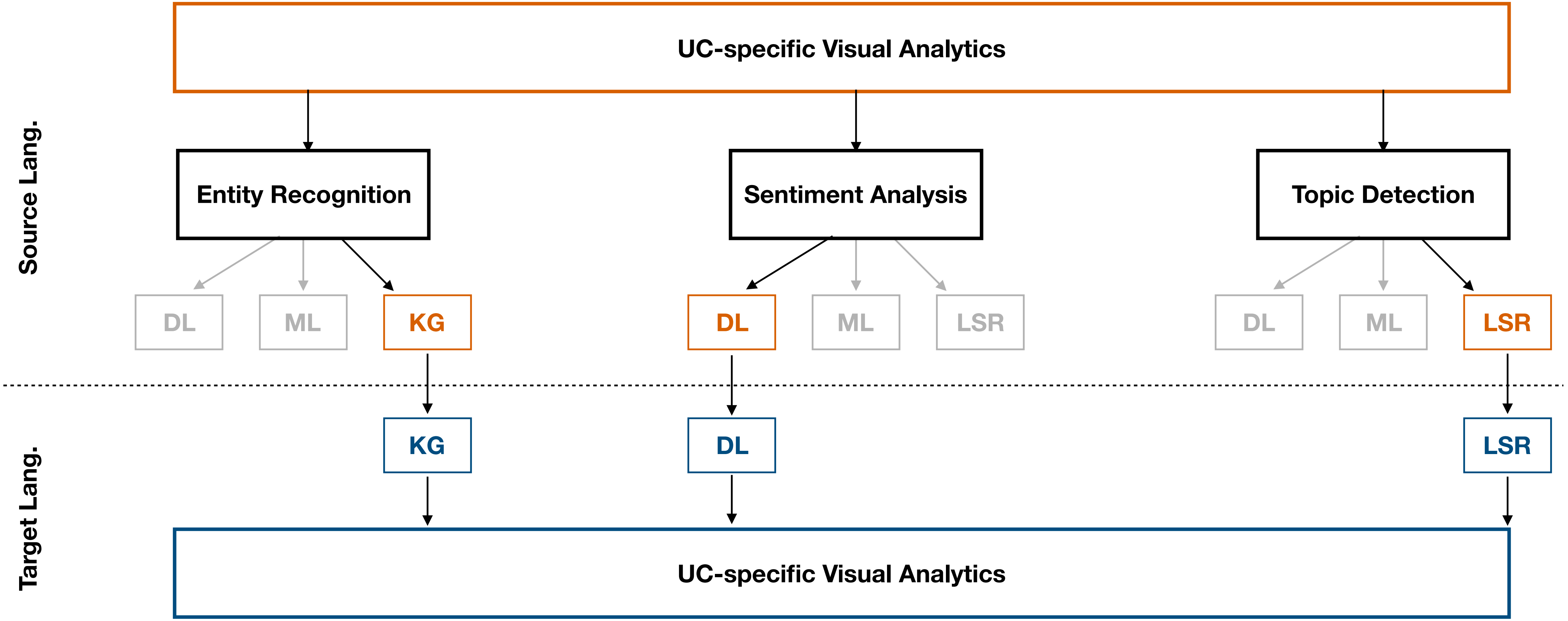
Case Study II - Overview

Example Use Case: „What is cardiologists’ perception on topic X for drug Y?“



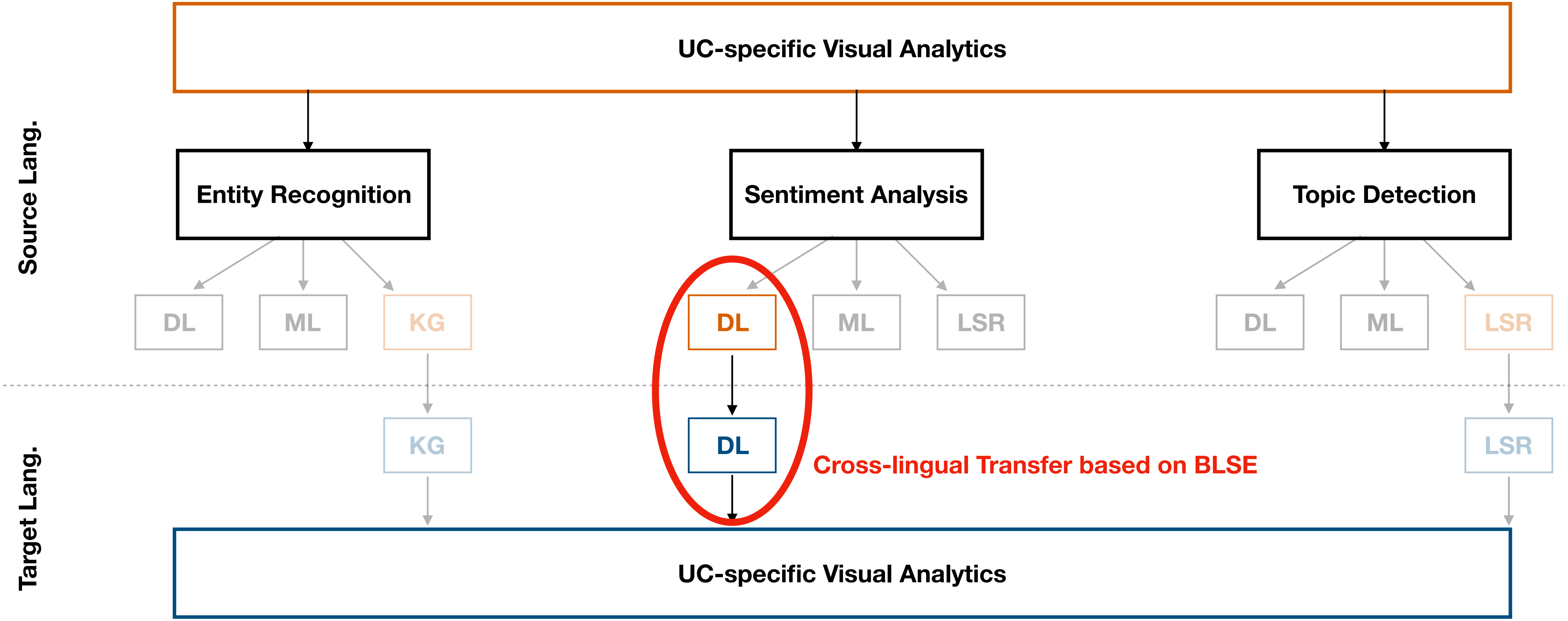
Case Study II - Overview

Example Use Case: „What is cardiologists’ perception on topic X for drug Y?“



Case Study II - Overview

Example Use Case: „What is cardiologists' perception on topic X for drug Y?“

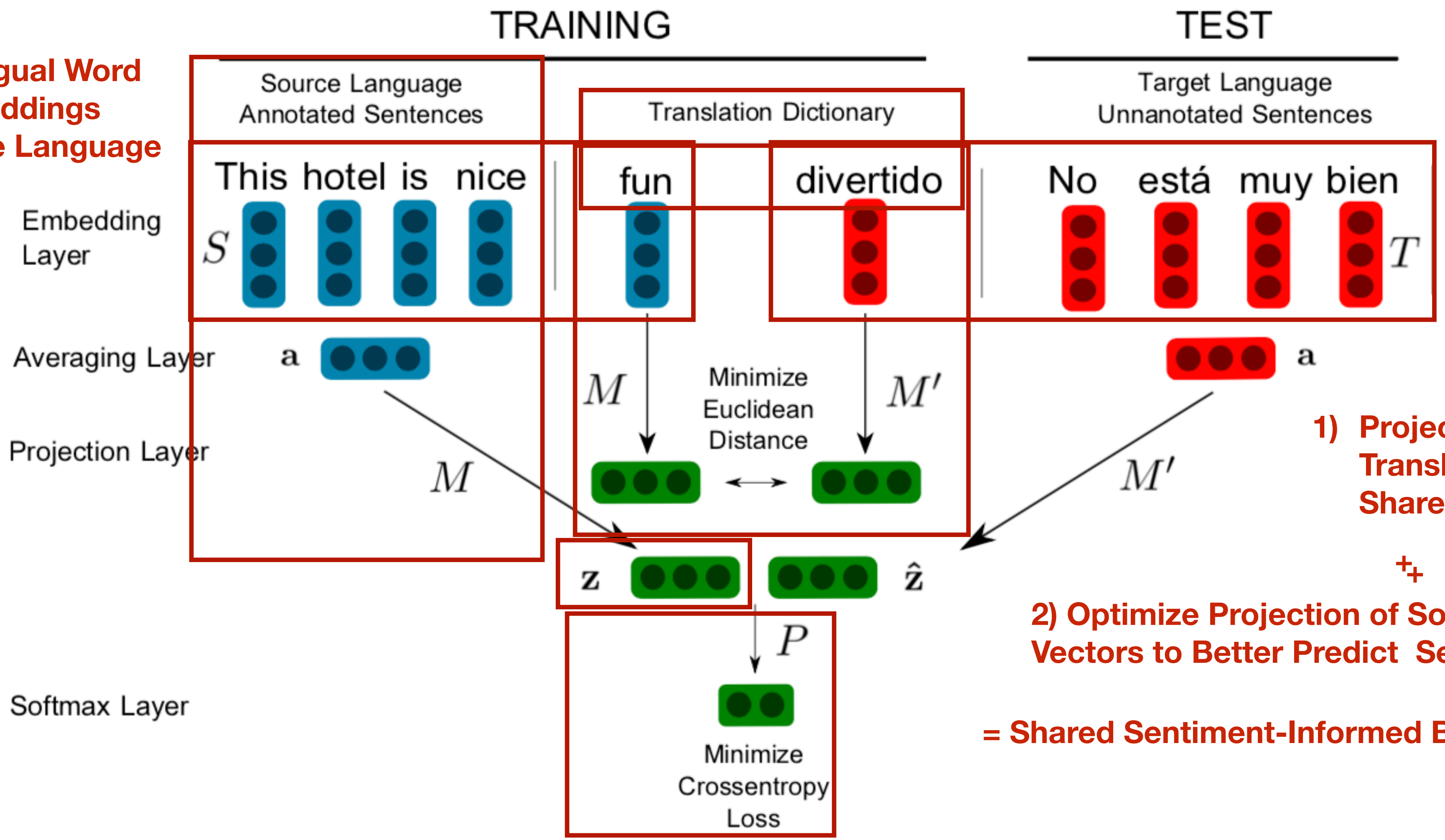


Cross-lingual Transfer based on BLSE

Case Study II - Model Architecture

Bilingual Dictionary for Source-Target Language Pair

Monolingual Word Embeddings for Source Language



Monolingual Word Embeddings for Target Language

1) Project Dictionary Translation Pairs Close in Shared Bilingual Space

+

2) Optimize Projection of Source Sentence Vectors to Better Predict Sentiment Labels

= Shared Sentiment-Informed Bilingual Space

Source: Barnes et al. 2018

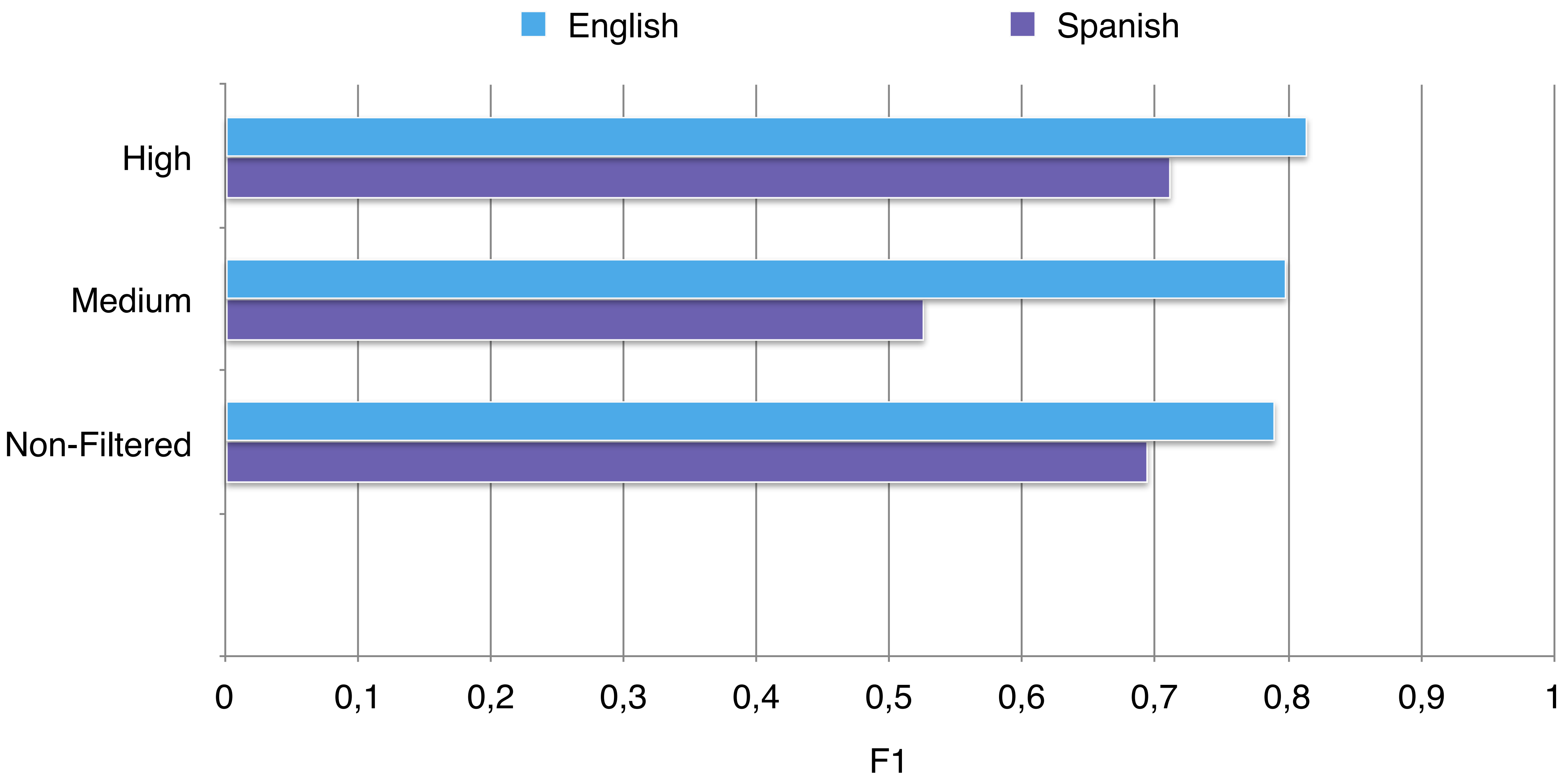
Case Study II - Preliminary Experiments

1. **Coverage of Embedding Vocabulary** - We expect pre-trained embeddings to not match our type of text well. What impact on performance has the degree of overlap of samples with the word embeddings' vocabulary?

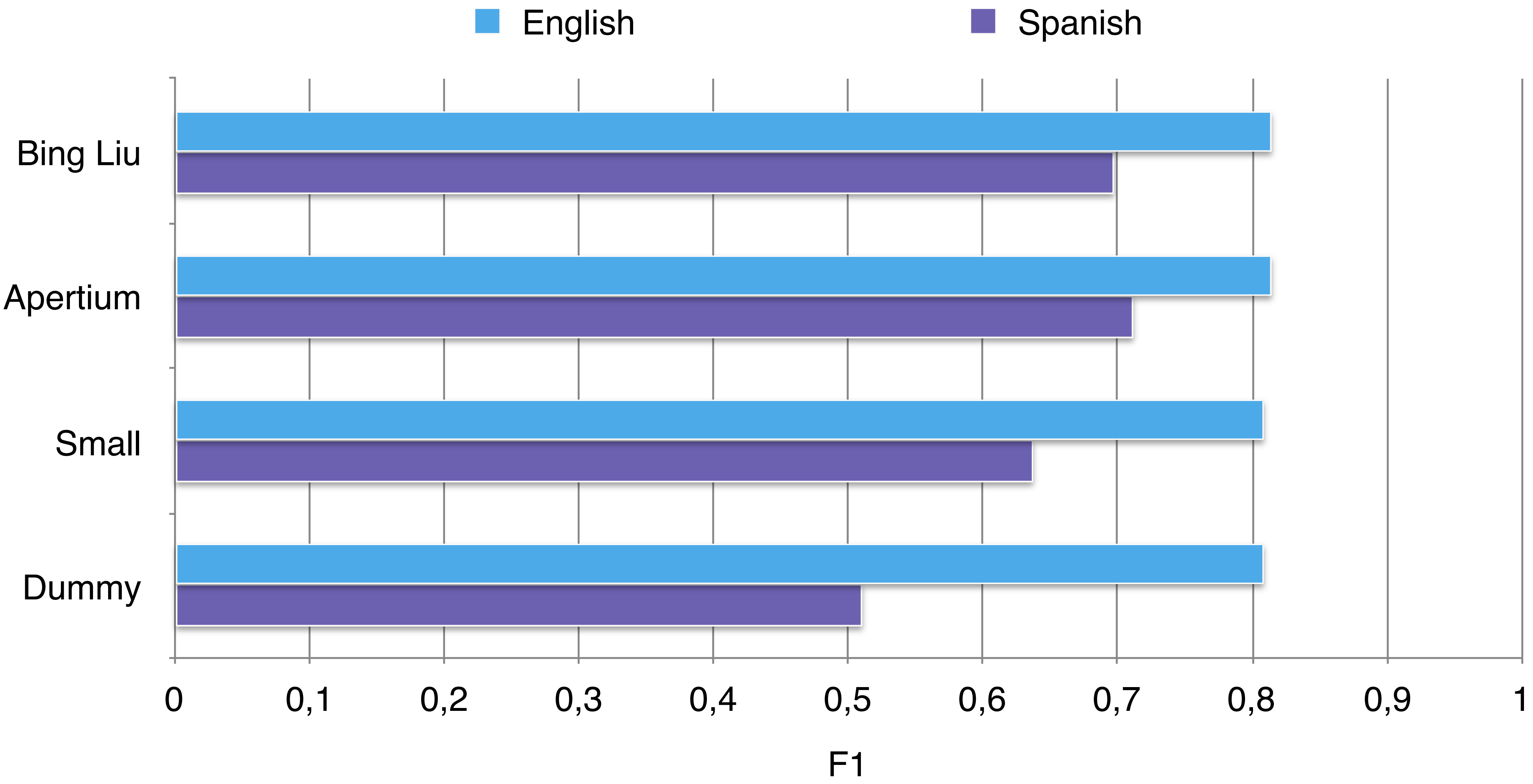
2. **Bilingual Lexica** - How do different lexica impact performance?

3. **Domain of Word Embeddings** - How does performance differ for embeddings trained on news vs domain-specific (biomedical) corpora?

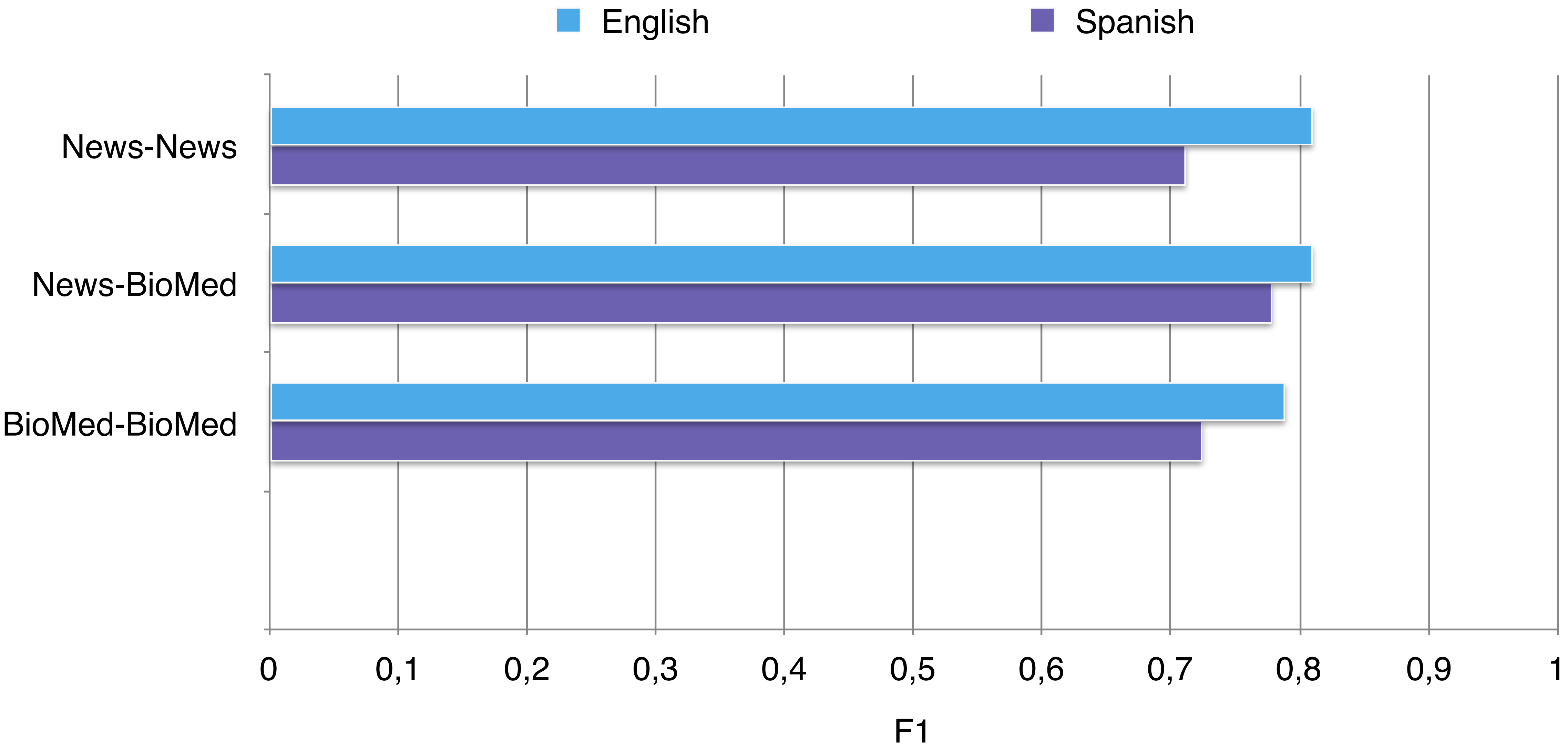
Case Study II - Coverage of Embeddings Vocab



Case Study II - Bilingual Lexicon



Case Study II - Domain of Word Embeddings



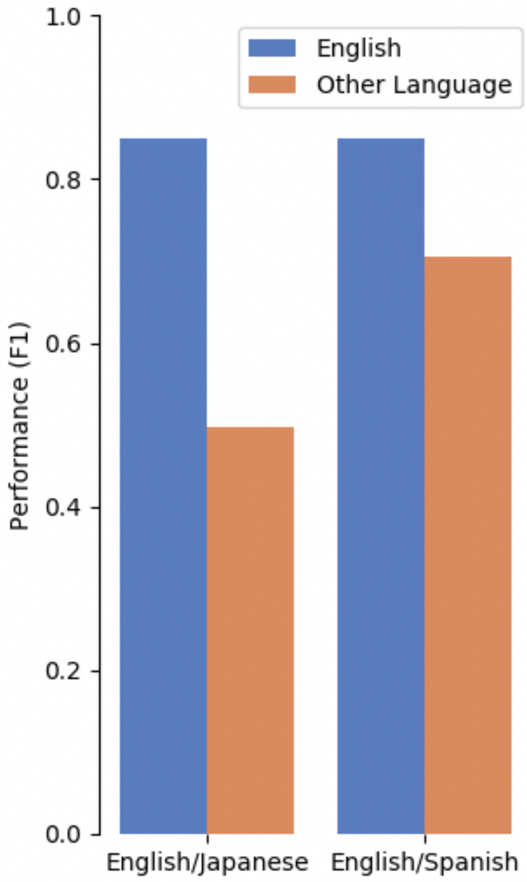
Outline

1. Background and Motivation
2. Case Study I: Machine Translation Pipeline (English-Japanese)
3. Case Study II: Cross-Lingual Transfer (English-Spanish)
4. **Conclusion**

Conclusion

Multilingual analytics based on Machine Translation vs Cross-Lingual Transfer

- Machine Translation enables multilingual analytics with low effort, but has inherent limitations that are not easily mitigated in a robust way
- Cross-lingual transfer is more involved and depends on the availability of language resources (LLOD helps!), but these are also an opportunity for adaption to task and domain. With an optimised choice of resources, we get close to source-language-level performance in preliminary experiments!
- Results can hardly be compared across case studies (different language pairs, different samples, different architectures) - however, first results from cross-lingual transfer are promising. Is it better in our specific setting? Principled comparison to come!



Thank you for your attention!

Matthias Orlikowski, NLP Product/Solutions Engineer

matthias.orlikowski@semalytix.de

Susana Veríssimo, Junior Language Engineer

susana.verissimo@semalytix.de

Matthias Hartung, Chief Research & Development Officer

hartung@semalytix.de

| Semalytix GmbH | Meller Straße 2 | 33613 Bielefeld | Germany |

SEMALYTIX

Prêt-à-LLOD

SEMALYTIX