Beyond Automatic Translation: Aligning Wikipedia Sections Across Multiple Languages

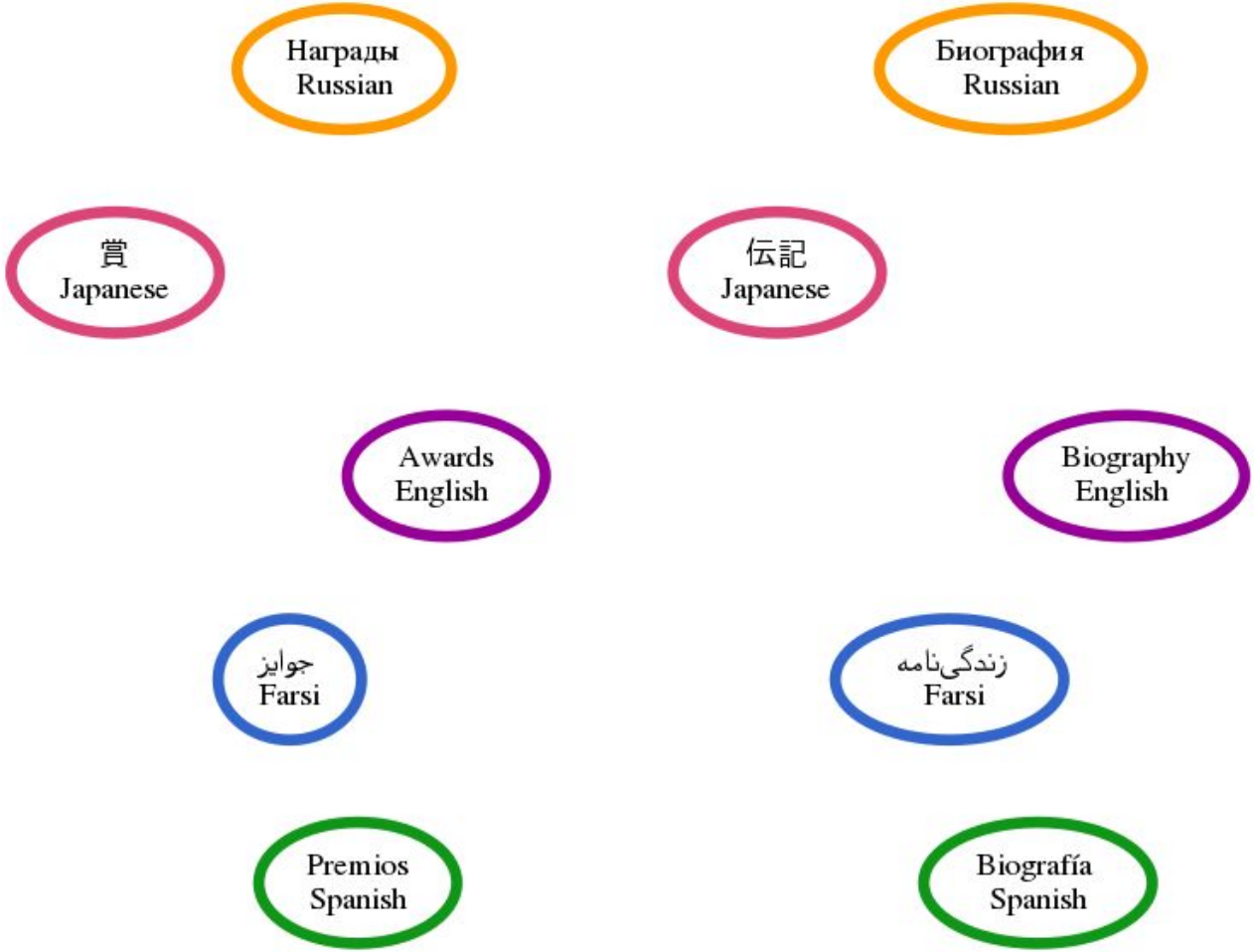
Bahodir Mansurov, Diego Sáez-Trumper, Robert West, Leila Zia

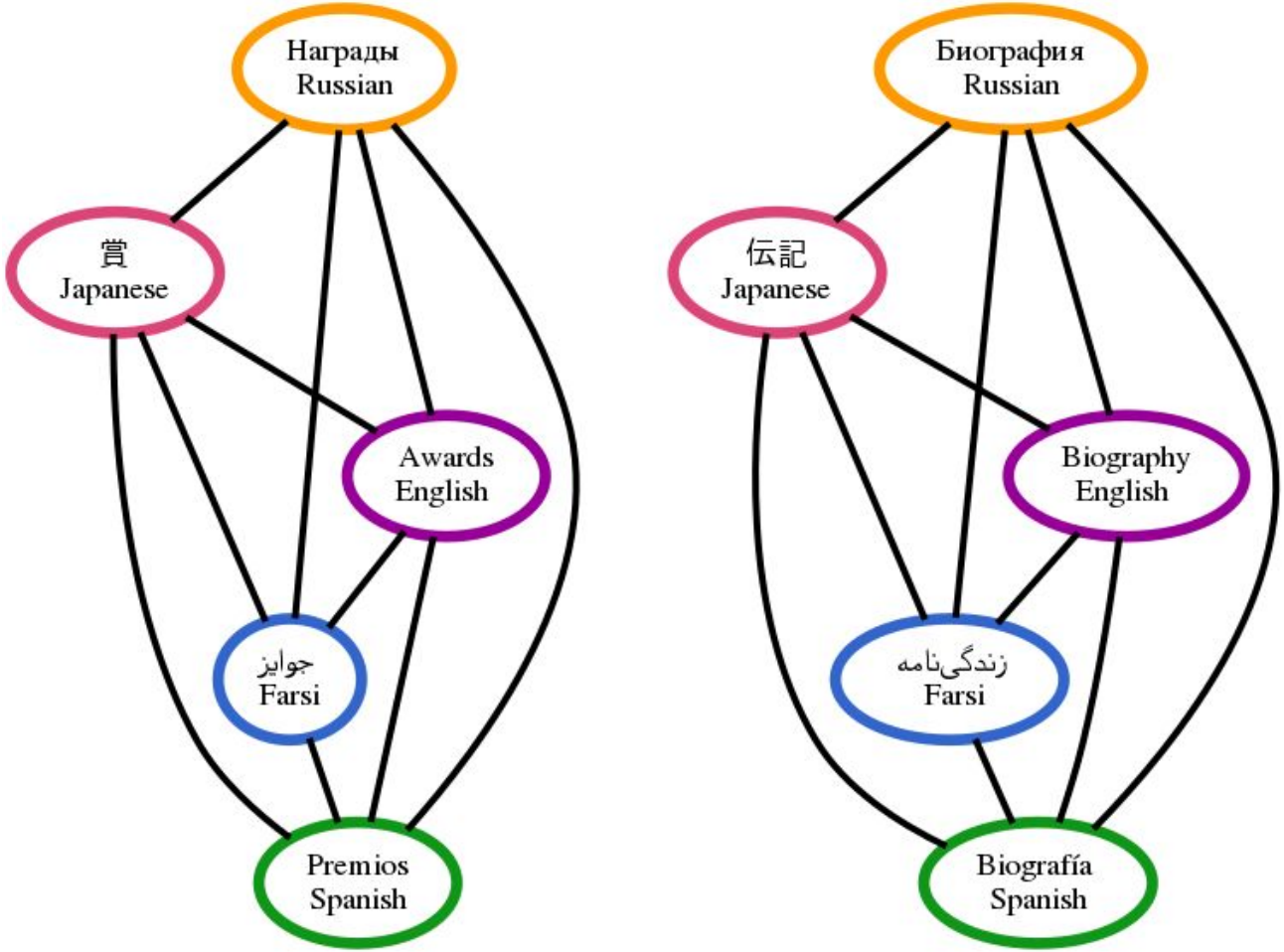
Research Showcase
2018-03-21

Create a data set with section names aligned across languages

Why?

- Cross-lingual section recommendations.
- Improve the content translation tools.
- Towards generate an abstract ontology for section titles.


Награды
Russian
Биография
Russian
賞
Japanese
伝記
Japanese
Awards
English
Biography
English
جوایز
Farsi
زندگی‌نامه
Farsi
Premios
Spanish
Biografía
Spanish


Награды
Russian
賞
Japanese
Awards
English
جوایز
Farsi
Premios
Spanish
Биография
Russian
伝記
Japanese
Biography
English
زندگی‌نامه
Farsi
Biografía
Spanish

Can we trust Automatic Translation services?

Can we trust Automatic Translation services?

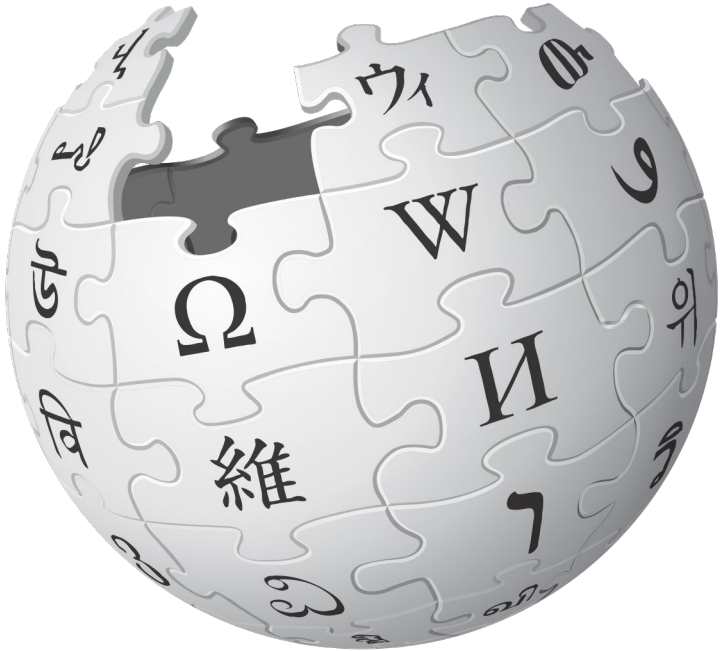
Accuracy depends on language
(ex. we found very different results for English to Spanish, and English to Farsi)

Challenges & Constraints

- Keep the style and conventions of each wikipedia edition.
- Equal importance to all languages
- Ground truth is difficult to build (e.g. Latvian to Bengali)

Assets & Opportunities

- Large and active community.
- Self reported language skills (Babel Template).
- Entities links across languages
- Large set of pre-trained models using Wikipedia (e.g. babylon project).



Building a training dataset

- Select a set of diverse languages:
  - Different scripts
  - Different families

- اللغة العربية المعيارية الحديثة
- **Русский язык**
- **Français**
- **Español**
- **English**
- **日本語**

Building a training dataset

- Select a set of diverse languages:
  - Different scripts
  - Different families

- **Modern Arabic**
- **Russian**
- **French**
- **Spanish**
- **English**
- **Japanese**

Wikipedia:Babel

en-5	This user is able to contribute with a **professional** level of **English**.
sv	Den här användaren talar **svenska** som **modersmål**.
no-4	Denne brukaren/brukeren meistrar/behersker **norsk** på **morsmålsnivå**.
he-3	משתמש זה מסוגל לתרום ברמה **מתקדמת** של **עברית**.
lt-2	Šis vartotojas gali prisidėti prie projekto **vidutinio lygio lietuvių** kalba.
es-1	Este usuario puede contribuir con un nivel **básico** de **español**.
an-0	Iste usuario **no repleca** l'**aragonés** (u el repleca con prou dificultat).

```
SELECT user.user_name,
       babel.babel_lang AS lang,
       babel.babel_level AS level
FROM babel
LEFT JOIN user
     ON user.user_id = babel.babel_user
ORDER BY user.user_name ASC;
```

https://quarry.wmflabs.org

Wikipedia:Babel

en-5	This user is able to contribute with a **professional** level of **English**.
sv	Den här användaren talar **svenska** som **modersmål**.
no-4	Denne brukaren/brukeren meistrar/behersker **norsk** på **morsmålsnivå**.
he-3	משתמש זה מסוגל לתרום ברמה **מתקדמת** של **עברית**.
lt-2	Šis vartotojas gali prisidėti prie projekto **vidutinio lygio lietuvių** kalba.
es-1	Este usuario puede contribuir con un nivel **básico** de **español**.
an-0	Iste usuario **no repleca** l'**aragonés** (u el repleca con prou dificultat).

```
SELECT us
       ba
       ba
FROM babe
LEFT JOIN
     ON u
ORDER BY
```

```
ng,
evel


babel_user
```

https:/ .org

language pair	# of users
ar-en	139
ar-es	43
ar-fr	113
ar-ja	8
ar-ru	42
en-es	5991
en-fr	2796
en-ja	203
en-ru	5324
es-fr	2796
es-ja	48
es-ru	236
fr-ja	43
fr-ru	399
ja-ru	55
total	**18236**

Problem Definition

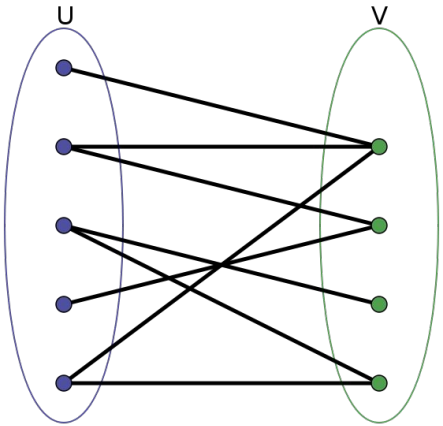
- Determine for each pair of sections (S1, S2), where S1 and S2 are from different languages, whether S1 is the translation of S2, across multiple languages.

Problem Definition

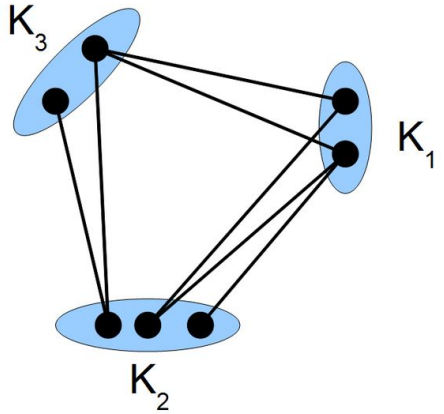
- Determine for each pair of sections (S1, S2), where S1 and S2 are from different languages, whether S1 is the translation of S2, across multiple languages.
- Consider a link prediction task, in multipartite (k-partite) graph, where each k group correspond to a language, nodes are section titles, and links represent translation.

Multipartite (k-partite) Graph

- k-partite graph is a graph whose vertices are or can be partitioned into k different independent sets.
- k=2 -> bipartite graph (ex. Actors and Movies)
- In our case k is equal to the number in languages in our dataset.



K = 2



K = 3

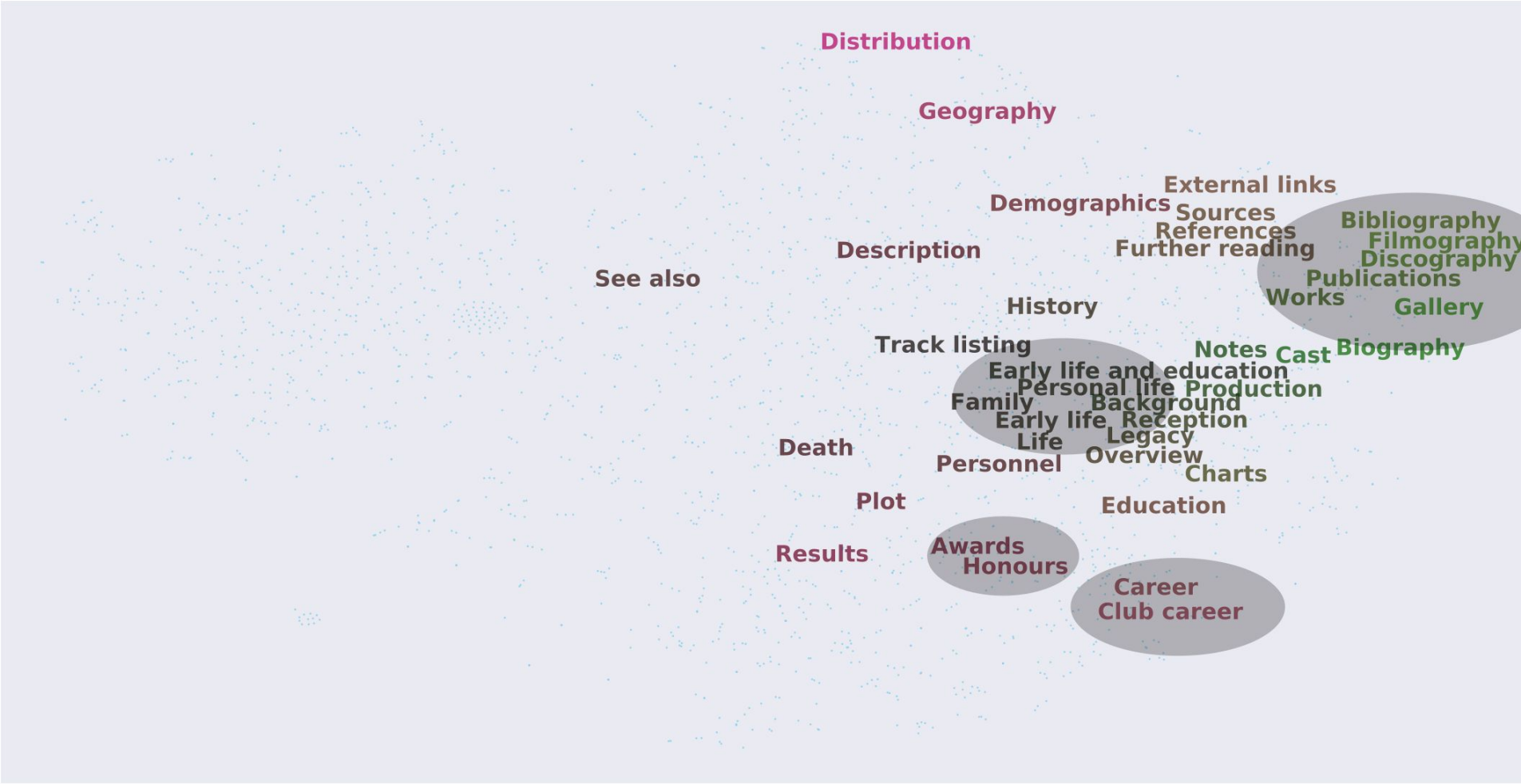
Features

- Automatic Translation
- Levenshtein distance
- Outcoming links
- Word embeddings
- Co-occurrence counts

Embeddings

"When some object X is said to be embedded in another object Y, the embedding is given by some injective and structure-preserving map f : X → Y”

https://en.wikipedia.org/wiki/Embedding

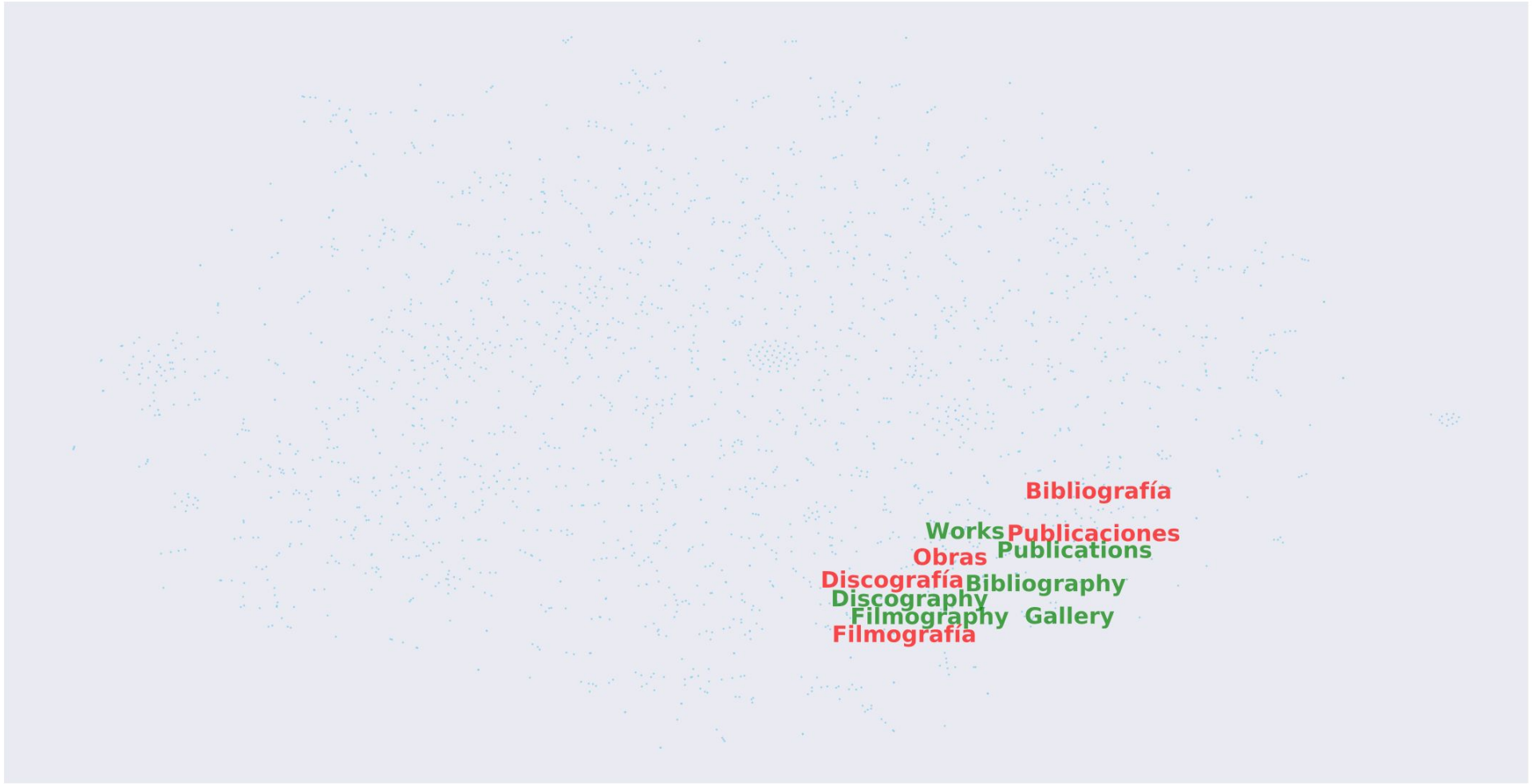

Distribution
Geography
External links
Demographics
Sources
Bibliography
References
Filmography
Description
Further reading
Discography
See also
Publications
Works
History
Gallery
Track listing
Notes
Cast
Biography
Early life and education
Personal life
Production
Family
Background
Early life
Reception
Life
Legacy
Death
Overview
Personnel
Charts
Plot
Education
Results
Awards
Honours
Career
Club career


Filmografía
Discografía
Publicaciones
Bibliografía
Bibliography
Filmography
Discography
Publications
Works

Aligned word embeddings

- Babylon Project:
  - Dictionary based alignment across languages
  - Linear transformation

https://github.com/Babylonpartners/fastText_multilingual

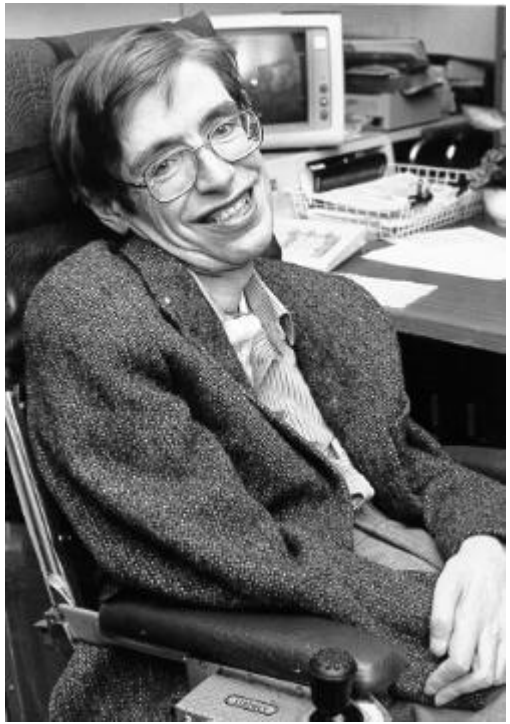

Bibliografía
Works
Publicaciones
Publications
Obras
Discografía
Bibliography
Discography
Filmography
Gallery
Filmografía

Co-occurrence counts

- We use Wikidata to align articles, counting the co-occurrence of sections in articles about the same item across languages

Contents [hide]

Stephen Hawking (Q17714)



Índice [ocultar]

Please help us to create the training dataset

[[m:Research:Expanding_Wikipedia_articles_across_languages]]
T183039

Synonym Detection

Problem Definition

- Determine for each pair of sections (S1, S2) from the same language whether they are synonyms of each other.

Features

- Word embeddings
- Levenshtein distance
- IsSubset
- Tf-Idf similarity (*)

Synonym Detection

- **Intuition**: Two sections are likely to be synonym if both tends to co-occur with similar sections, but (almost) never co-occurs among themselves.

Synonym Detection

- **Intuition**: Two sections are likely to be synonym if both tends to co-occur with similar sections, but (almost) never co-occurs among themselves.
- **Problem**: Most frequent sections tends to co-occur with almost all the sections.

Synonym Detection

- **Solution**:
  - Represent sections (S) as TF-IDF vectors, using the counting of co-occurrences as TF.
  - Measure cosine similarity of TF-IDF vectors, between pairs of sections S1 and S2
  - Divide by the number of co-occurrences (+1) between S1 and S2

Examples

S1	S2	tfIdfSimilarity
Stud career	Stud record	0.99
Side effects	Adverse effects	0.99
Musical career	Music career	0.98
Home video	Home media	0.97
Line-up	Band members	0.96

Future Work & Tasks

- Gather labels for section alignment and synonyms.
- Test frameworks for the link detection.
- Improve word embedding alignments.

References

- K-partite graphs: https://en.wikipedia.org/wiki/Multipartite_graph
- Embeddings: https://en.wikipedia.org/wiki/Embedding
- Word Embeddings: https://en.wikipedia.org/wiki/Word_embedding
- Babylon Project: https://github.com/Babylonpartners/fastText_multilingual
- Code: https://github.com/digitalTranshumant/wmf-interlanguage

Questions?