CLRL: Feature Engineering for Cross-Language Record Linkage

Öyku Özlem Çakal
Technische Universität Berlin
o.cakal@campus.tu-berlin.de

Mohammad Mahdavi
Technische Universität Berlin
mahdavilahijani@tu-berlin.de

Ziawasch Abedjan
Technische Universität Berlin
abedjan@tu-berlin.de

P R O B L E M

Research Question

- How to link the records in a cross-language setting?
  - Where each input dataset is in a different language

Challenges

- Simple translation of datasets may not work
  - Ambiguity in translation
  - Out-of-vocabulary (OOV) terms

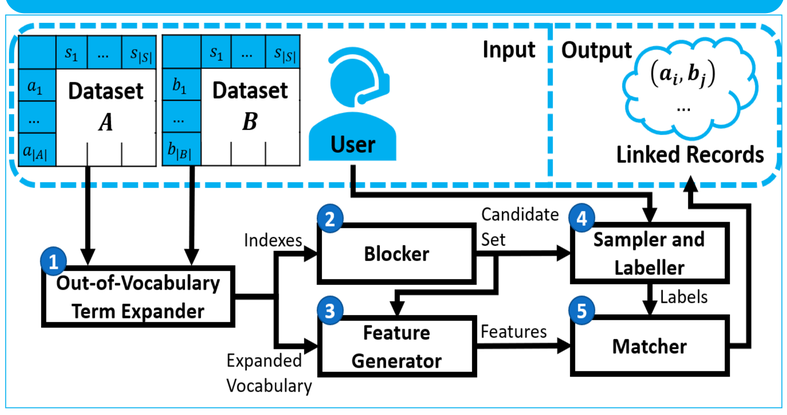
Motivation Example

- Linking the movie “Forbidden Planet” to its German equivalent “Alarm im Weltall” is not trivial
  - They have no lexical similarity
  - The translated title “Alarm in the Universe” is still not similar to “Forbidden Planet”

ID	Name	Year	ID	Name	Jahr
1	Heat	1995	1	Alarm im Weltall	1956
2	Forbidden Planet	1956	2	Der Pate	1972

S O L U T I O N

The Workflow of CLRL



OOV Term Expansion

- Morphological checking
  - E.g., compound word “firstsight”
- Spell checking
  - E.g., typo “researcch”
- Ostrich policy
  - E.g., named entity “Lebowski”

Feature Generation

- Monolingual similarity features
  - Typical lexical measures
- Multilingual similarity features
  - Via cross-language word embedding models [4]

Cross-Language Word Embedding Models…

- … map words of different languages into a shared multidimensional space
- … assign similar vectors to cross-lingual similar words, such as “dog” and “hund”

Multilingual Similarity Features

Mean Vector Similarity:
Pick the similarity of mean vectors for two data cells

Max Vector Similarity:
Pick the similarity of the most similar word pair for two data cells

Optimal Alignment Similarity:
Pick the mean similarity of word pairs when the best 1-to-1 word alignment is computed for two data cells

Max Alignment Similarity:
Pick the mean similarity of word pairs when each word is aligned to its most similar word in the other data cell

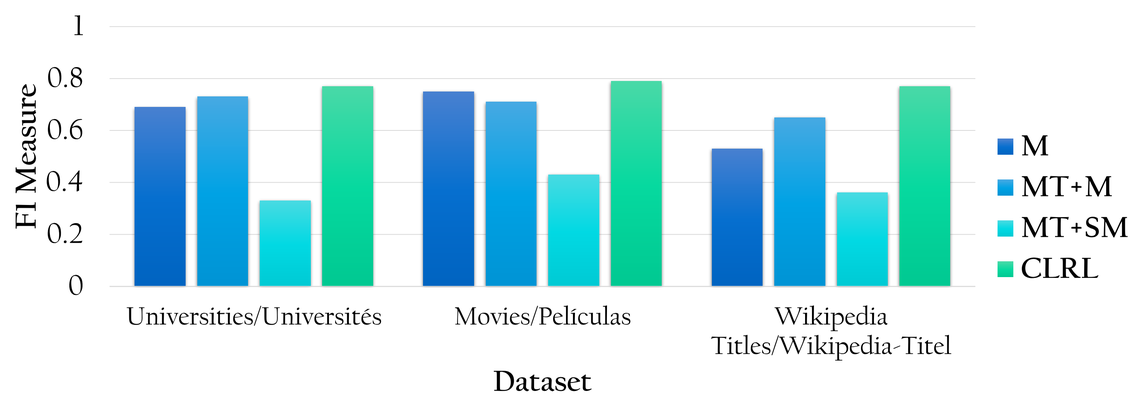
E V A L U A T I O N

Experimental Setup

- 3 Baselines
  - Magellan [1] (M)
  - Machine translation [2] + Magellan [1] (MT+M)
  - Machine translation [2] + semantic matching [3] (MT+SM)
- 6 Datasets

Name	Language	#Rows	#Common Attributes	#Actual Linked Records
Universities	English	8758	16	940
Universités	French	3957		
Movies	English	1273	14	72
Películas	Spanish	15334		
Wikipedia Titles	English	1976	2	83
Wikipedia-Titel	German	2159		

Effectiveness of CLRL



References

[1] Pradap Konda et al. 2016. Magellan: Toward building entity matching management systems. PVLDB 9, 12, 1197–1208.
[2] Zhifei Li et al. 2009. Joshua: An open source toolkit for parsing-based machine translation. In StatMT. 135–139.
[3] Yuting Song et al. 2016. Cross-language record linkage using word embedding driven metadata similarity measurement.
[4] Sebastian Ruder et al. 2017. A survey of cross-lingual word embedding models. arXiv preprint arXiv:1706.04902.

Source Code

Our prototype is available online: https://github.com/bigdama/clrl

Acknowledgement

This work has been partially funded by the German Ministry for Education and Research as BBDC II (01IS18025A).

Bundesministerium für Bildung und Forschung