UBC_UOS-TYPED: Regression for Typed-similarity

Eneko Agirre, Aitor Gonzalez-Airre, German Rigau
IXA NLP Group, UPV/EHU
{e.agirre, aitor.gonzalez-agirre, german.rigau}@ehu.es

Nikolaos Aletras, Mark Stevenson
University of Sheffield
{n.aletras, m.stevenson}@dcs.shef.ac.uk

GOAL AND APPROACH

The typed-similarity pilot task attempts to characterize the ***reason*** and/or ***type*** of similarity. We used a range of heuristics that rely on information from the appropriate meta-data fields for each type of similarity and we trained a linear regressor.

The typed-similarity dataset

The dataset derives from pairs of Cultural Heritage items from Europeana. The items comprise meta-data describing a cultural heritage item, including a thumbnail of the item.

In addition to general similarity, the dataset includes specific kinds of similarity: **author**, **people involved**, **time period**, **location**, **event or action**, **subject** and **description**.

Basic system

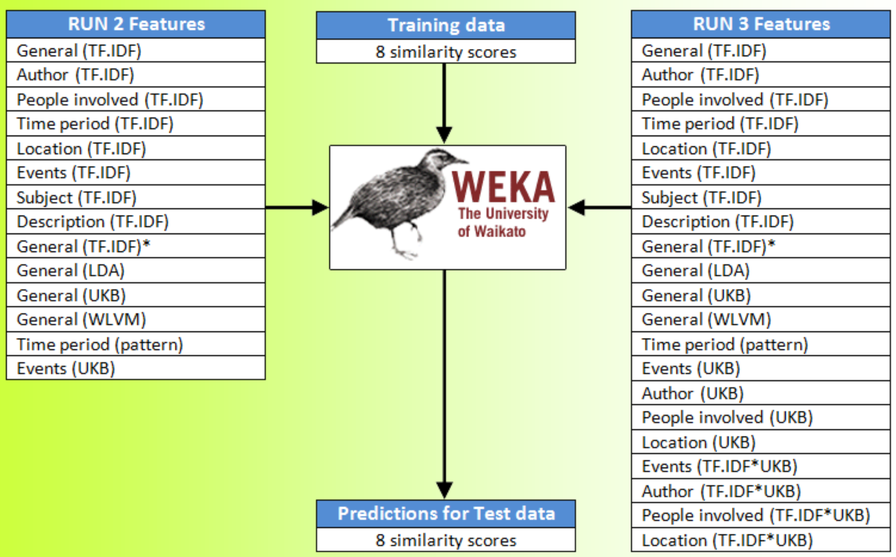
We analyzed the text in the metadata, performing *lemmatization*, *PoS tagging*, *named entity recognition and classification* (*NERC*) and *date detection* using **Stanford Core NLP**.

- **General**: *cosine similarity* of **TF.IDF** vectors of tokens, taken from all fields.
- **Author**: *cosine similarity* of **TF.IDF** vectors of dc:Creator field.
- **People involved**, **time period** and **location**: *cosine similarity* of **TF.IDF** vectors of location/date/people entities recognized by *NERC* in all fields.
- **Events**: *cosine similarity* of **TF.IDF** vectors of event *verbs* and *nouns* in all fields. A list of verbs and nouns possibly denoting events was derived using the *WordNet Morphosemantic Database*.
- **Subject** and **description**: *cosine similarity* of **TF.IDF** vectors of respective fields.

Machine Learning: Linear regressor

We take the indicator from the basic systems as features, and use linear regression (as made available by **Weka**) to learn models that fit the training data.

We generated further similarity scores for general similarity, including *Latent Dirichlet Allocation* (**LDA**), *Personalized PageRank* (**UKB)** and *Wikipedia Link Vector Model* (**WLVM**) using information taken from all fields.



We generated 8 models, one for each of the similarity types.

Results

On training data:

Run	General	Author	People	Time	Locat.	Event	Subj.	Desc.	Mean
1	.7269	.4474	.4648	.5884	.4801	.2522	.4976	.5389	.5389
2	.7777	.6680	.6767	.7609	.7329	.6412	.7516	.8024	.8024
3	**.7866**	**.6941**	**.6965**	**.7654**	**.7492**	**.6551**	**.7586**	**.8067**	**.8067**

On test data:

Run	General	Author	People	Time	Locat.	Event	Subj.	Desc.	Mean	Rank
1	.7256	.4568	.4467	.5762	.4858	.3090	.5015	.5810	.5103	6
2	.7457	.6618	.6518	**.7466**	.7244	.6533	.7404	.7751	.7124	4
3	**.7461**	**.6656**	**.6544**	.7411	**.7257**	**.6545**	**.7417**	**.7763**	**.7132**	**3**

Conclusions

- We combined some simple heuristics for each similarity, based on appropriate metadata fields.
- The use of lineal regression improved the results considerably across all types.
- Our team ranked second in the competition.