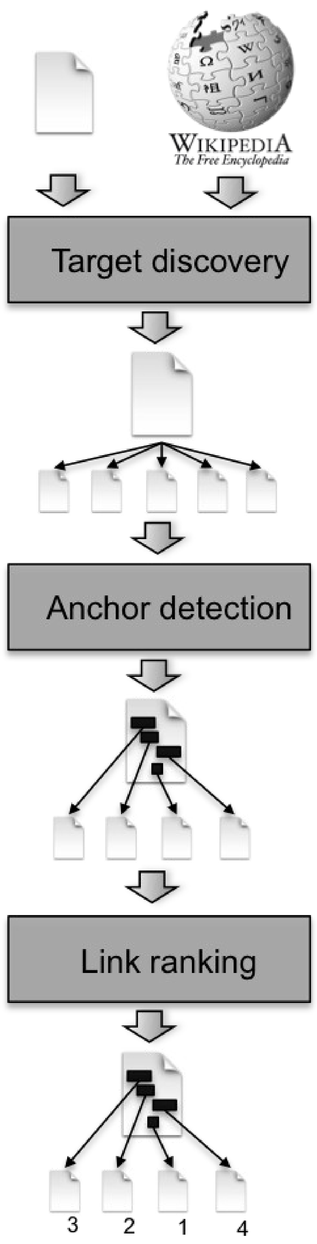
KMI, The Open University at NTCIR-9 CrossLink: Cross-Lingual Link Discovery in Wikipedia Using Explicit Semantic Analysis

Petr Knoth, Lukas Zilka, Zdenek Zdrahal

The CrossLink task (Cross-Lingual Link Discovery - CLLD) is a way of automatically finding links between documents in different languages. We present Cross-Lingual Link Discovery (CLLD) methods that utilise Explicit Semantic Analysis to suggest a set of cross-lingual links from an English Wikipedia article to articles in another language.



Target discovery

Takes as an input a new "orphan" document (i.e. a document that is not linked to other documents) written in the source language and automatically generates a list of potential target documents.

Two approaches

ESA2Links – consists of two steps. In the first step, ESA vectors are calculated for each document in the document collection and also for the orphan document. Similarity between the resulting ESA vectors is then calculated and *k* most similar pages are identified. In the second step, the *k* most similar pages are taken as a seed and the system extracts from them all links in the form *[anchor, pageID]*. Using the cross-lingual mapping between Wikipedia pages, the *pageID*, describing a page in the source language, is mapped to an appropriate ID describing the same page in the target language.

Terminology – recommends as targets all pairs *[pageTitle, pageID]* in the whole Wikipedia for which there exists an explicit cross-lingual mapping between the source and the target language version of the page.

Anchor detection

Takes as an input the set of targets and tries to detect suitable anchors for them in the orphan document.

Link ranking

The approach used to rank the generated links is based on machine learning. Each link is first modeled by a set of features. The features are represented as a vector assuming their mutual independence. An SVM is then used to decide whether a link should be included and with which confidence.

Features:

- *ESA similarity* is a real number between 0 and 1, which expresses the similarity of texts. Three different features were included:
  - Similarity of the link text to the target document text.
  - Similarity of the link text to the target document title.
  - Similarity of the input document text to the target document text.

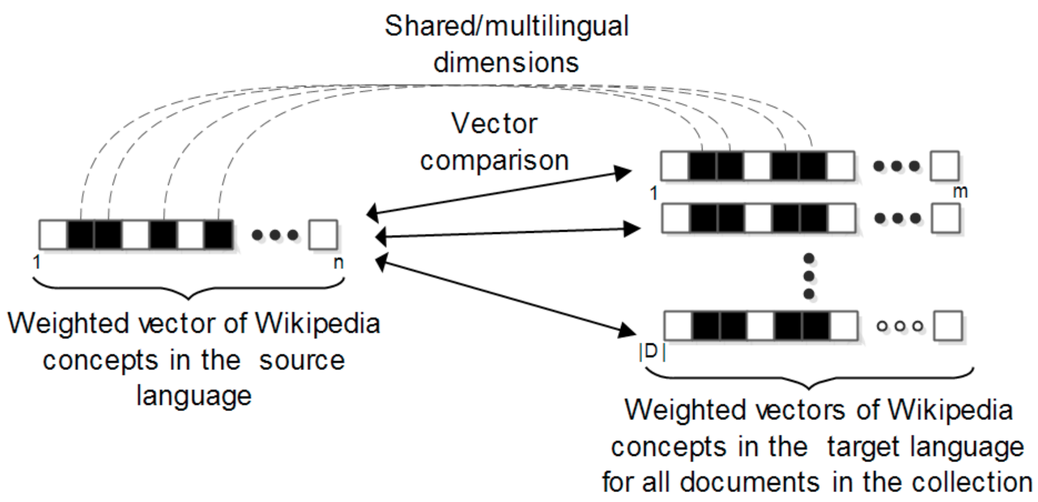
• *Generality* is a measure expressing how general a given topic is. It is an integer number between 0 and 16 defined as the minimum depth at which the topic is located in Wikipedia's category tree.

• *Link frequency* is a measure expressing how many times a particular keyword occurs as a link (or more precisely as an anchor) in the whole document collection.

• *Occurrence* of the link text in the input document is a relative measure of the first, last and current occurrence of the link text in the input document, and the difference between its first and last occurrence.

Cross-lingual discovery

ESA Discovery - We have also tested a method that does not directly rely on the manually created mappings between Wikipedia pages in the target discovery step. This method utilises Cross-Lingual Explicit Semantic Analysis (CL-ESA) (Sorg & Cimiano, 2008) to discover an equivalent page in the target language (Chinese) for a page in the source language (English) by measuring cross-language semantic similarity.



KMI runs

Four runs for NTCIR CrossLink for English to Chinese

Run 1: KMI_SVM_ESA_TERMDB - combines *ESA2Links* with *Terminology*

Run 2: KMI_SVM_ESA - applies *ESA2Links* for target discovery.

Run 3: KMI_SVM_TERMDB - uses *Terminology* only for target discovery.

Run 4: KMI_ESA_SVM_ESAdiscovery - uses *ESA2Links* for target discovery and *ESA discovery* for the cross-language step.

Results

The KMI methods scored first in Precision-at-5 in the A2F manual assessment and third in terms of R-Prec. Our methods were also second in F2F manual assessment in terms of MAP and R-Prec and third in terms of Precision-at-5. Our system ranked third in F2F Wikipedia ground-truth evaluation in terms of all MAP, R-Prec and Precision-at-5.

Run ID	MAP	R-Prec	P@5	P@10	P@20	P@30	P@50	P@250
F2F performance with Wikipedia ground truth								
KMI_SVM_ESA_TERMDB	**0.260**	**0.345**	0.712	**0.664**	0.530	0.491	**0.434**	**0.166**
KMI_SVM_ESA	0.251	0.338	**0.728**	**0.664**	**0.540**	**0.493**	0.430	0.153
KMI_SVM_TERMDB	0.127	0.211	0.624	0.552	0.454	0.383	0.302	0.078
KMI_ESA_SVM_ESADiscovery	0.059	0.148	0.264	0.240	0.186	0.165	0.138	0.044
F2F performance with manual assessment results								
KMI_SVM_ESA_TERMDB	**0.258**	**0.393**	0.720	**0.728**	**0.684**	0.648	0.604	**0.358**
KMI_SVM_ESA	0.231	0.344	0.728	0.720	0.678	**0.668**	**0.615**	0.306
KMI_SVM_TERMDB	0.133	0.192	**0.752**	0.692	0.636	0.613	0.561	0.178
KMI_ESA_SVM_ESADiscovery	0.054	0.132	0.464	0.388	0.348	0.321	0.283	0.119
A2F performance with manual assessment results								
KMI_SVM_ESA_TERMDB	**0.097**	**0.114**	**0.368**	**0.368**	**0.330**	0.303	0.269	**0.142**
KMI_SVM_ESA	0.080	0.092	0.360	0.364	**0.330**	0.299	0.260	0.113
KMI_SVM_TERMDB	0.070	0.075	0.376	**0.368**	0.324	**0.316**	**0.297**	0.096
KMI_ESA_SVM_ESADiscovery	0.014	0.035	0.088	0.108	0.110	0.108	0.090	0.045

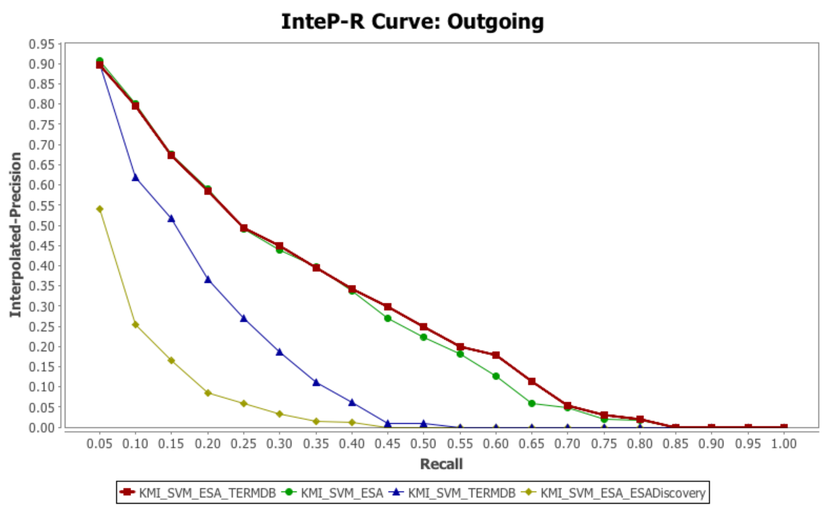


Figure 1. F2F performance using Wikipedia ground truth.

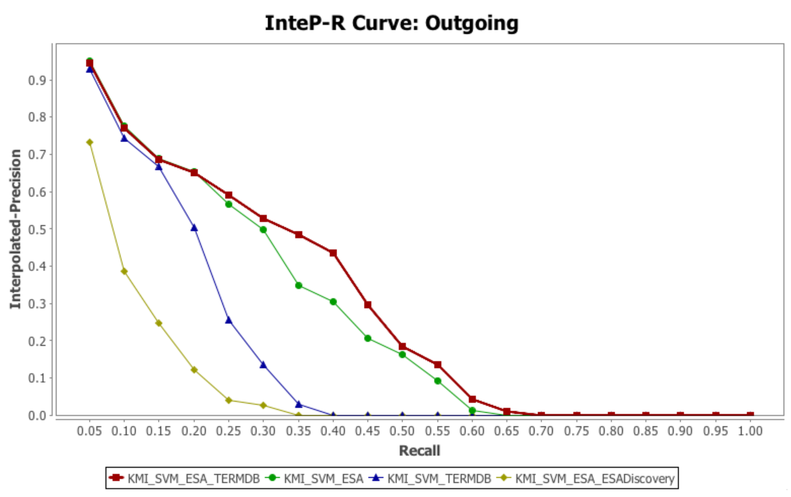


Figure 2. F2F performance of the KMI runs using manual assessment.

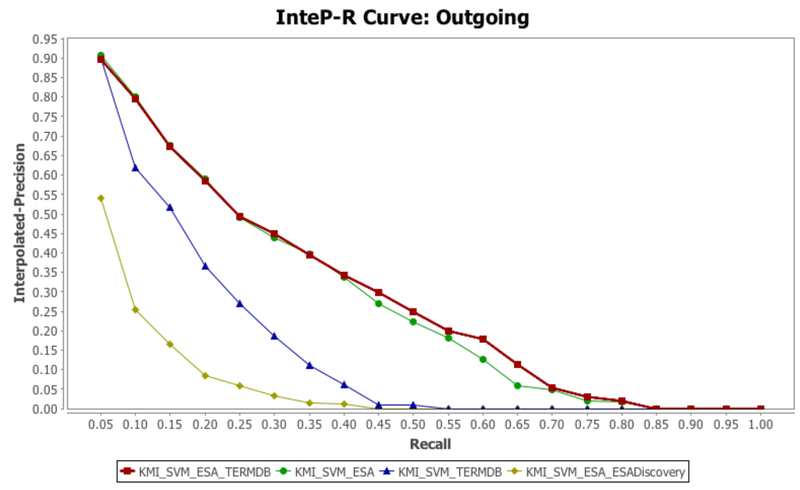


Figure 3. A2F performance using manual assessment.

Petr Knoth, Lukas Zilka, Zdenek Zdrahal
Knowledge Media Institute, The Open University{
{p.knoth, l.zilka, z.zdrahal}@open.ac.uk