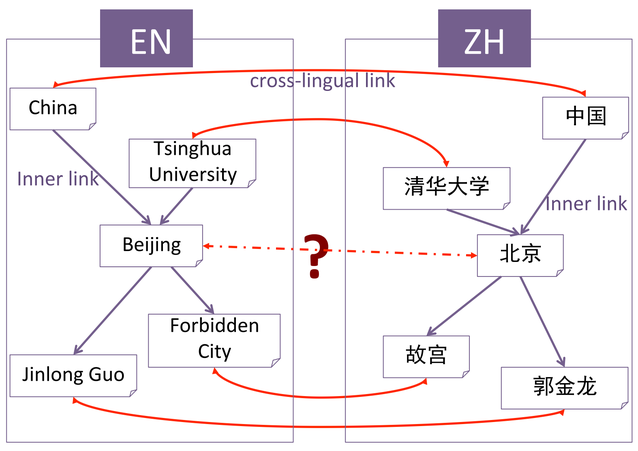
Boosting Cross-lingual Knowledge Linking via Concept Annotation

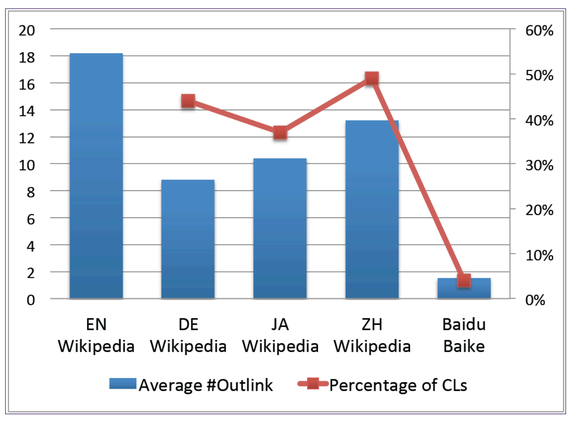
Zhichun Wang[1], Juanzi Li[2], Jie Tang[2]
[1]Beijing Normal University, [2]Tsinghua University

View this poster with the Poster in my Pocket app

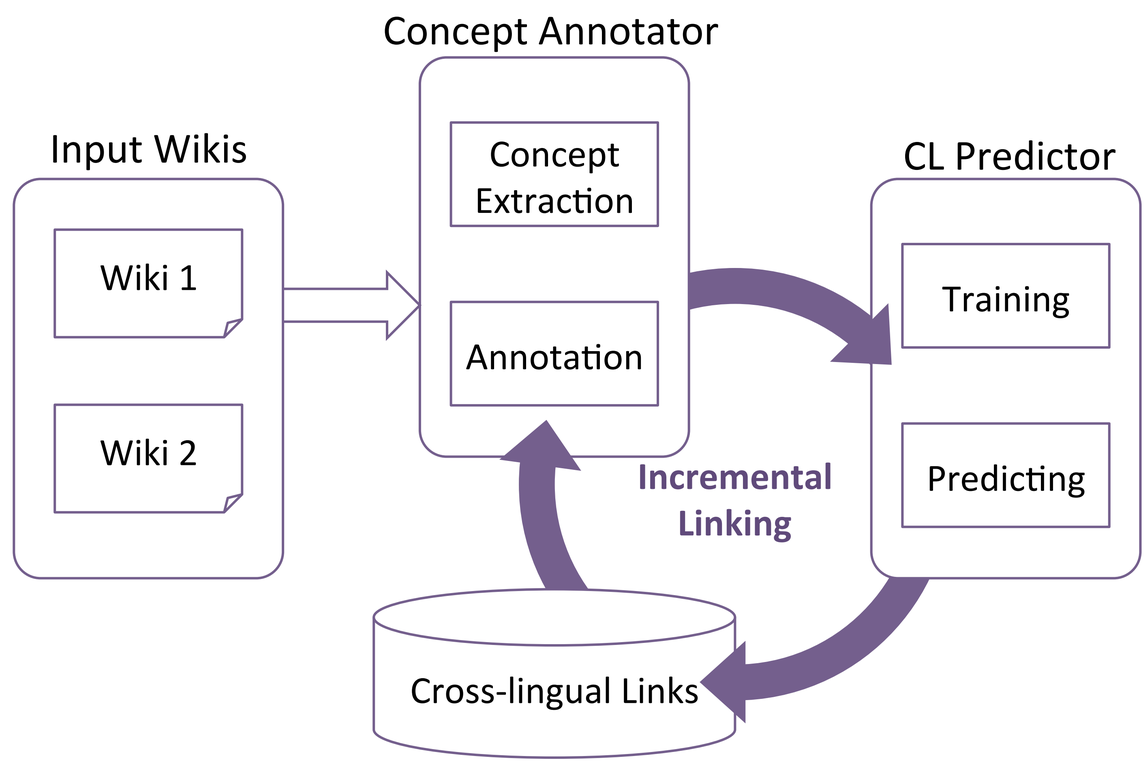
Motivation

Cross-lingual knowledge linking is to automatically discover cross-lingual links (CLs) between wikis, which can largely enrich the cross-lingual knowledge and facilitate sharing knowledge across different languages. The seed CLs and the inner link structures are two important factors for finding new CLs. **A challenging problem is how to find large scale CLs between wikis (e.g. English Wikipedia and Baidu Baike) when there are insufficient seed CLs and inner links?**

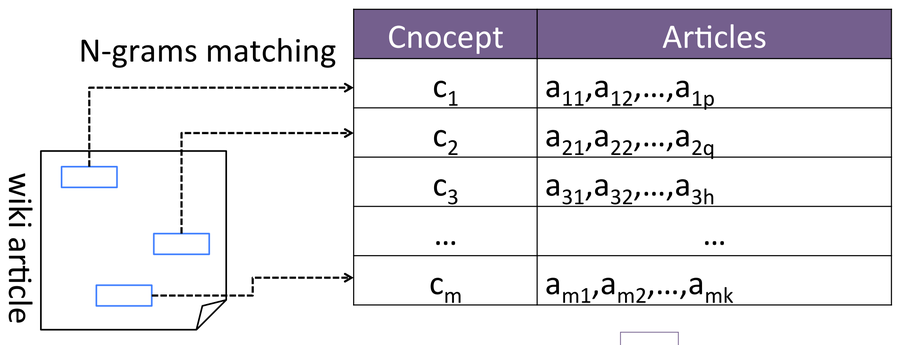




Proposed Framework



Annotating Concepts Within Each Wiki



Cnocept	Articles
c_1	$a_{11},a_{12},...,a_{1p}$
c_2	$a_{21},a_{22},...,a_{2q}$
c_3	$a_{31},a_{32},...,a_{3h}$
...	...
c_m	$a_{m1},a_{m2},...,a_{mk}$

Extract Concept-Article vocabulary

Annotation prediction

Link Probability

$$LP(a|c) = \frac{count(a,c)}{count(c)}$$

Semantic Relatedness

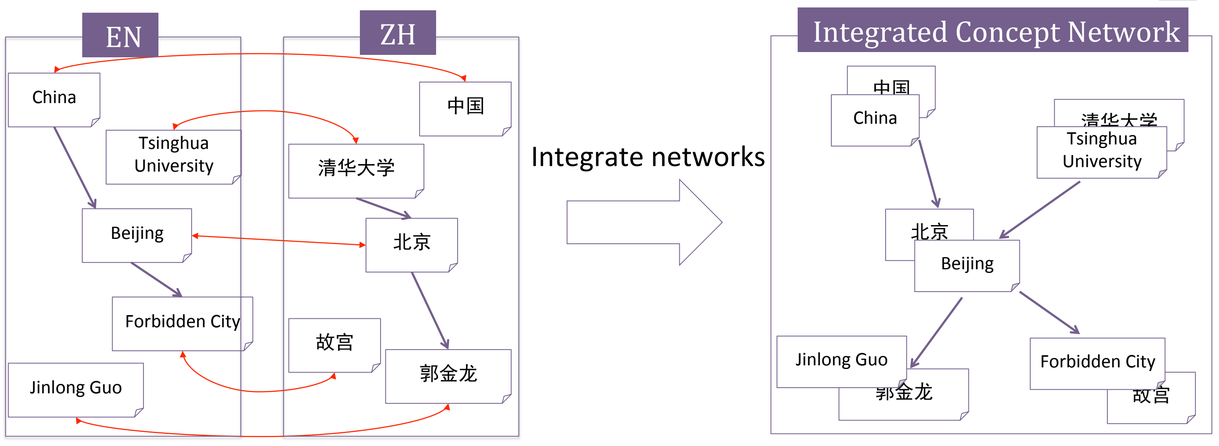
$$SR(a,c) = \frac{1}{|N_c|}\sum_{b\in N_c} r(a,b)$$

$$r(a,b) = 1 - \frac{log(max(|I_a|,|I_b|)) - log(|I_a \cap I_b|)}{log(|W|) - log(min(|I_a|,|I_b|))}$$

SR(a,c) is computed based on the links in the Integrated Concept Network

Annotation Criteria

$$b^* = argmax_{b\in T_{c_i}} LP(b|c_i) \times SR(b,c_i)$$



Predicting New Cross-lingual Links Between Wikis

Language-independent Features

Outlink similarity

$$f_1(a,b) = \frac{2\cdot|\phi_{1\to2}(O(a)) \cap O(b)|}{|\phi_{1\to2}(O(a))| + |O(b)|}$$

$$f_2(a,b) = \frac{2\cdot|\phi_{1\to2}(O^+(a)) \cap O^+(b)|}{|\phi_{1\to2}(O^+(a))| + |O^+(b)|}$$

Inlink similarity

$$f_3(a,b) = \frac{2\cdot|\phi_{1\to2}(I(a)) \cap I(b)|}{|\phi_{1\to2}(I(a))| + |I(b)|}$$

$$f_4(a,b) = \frac{2\cdot|\phi_{1\to2}(I^+(a)) \cap I^+(b)|}{|\phi_{1\to2}(I^+(a))| + |I^+(b)|}$$

Category similarity

$$f_5(a,b)(= \frac{2\cdot|\phi_{1\to2}(C(a)) \cap C(b)|}{|\phi_{1\to2}(C(a))| + |C(b)|}$$

$$f_6(a,b) = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\lambda(c_i^a, c_j^b)$$

Regression Model for Predicting New CLs

For each article *a* in W_1, the article *b** in W_2 that maximizes the following score function *s(a,b*)* and satisfies *s(a,b*)>0* is predicted as the corresponding article of *a*.

$$s(a,b) = \omega_0 + \vec{\omega}\cdot\vec{f}_{a,b}$$
$$= \omega_0 + \omega_1 \times f_1(a,b) + ... + \omega_6 \times f_6(a,b)$$

A regression model is used to learn the weights of features based on a set of known CLs. The optimal weights should satisfies:

$$\forall a_i \in A, \forall b' \in (B - \{b_i\}), s(a_i,b_i) - s(a_i,b') > 0$$

$$\vec{\omega}\cdot(\vec{f}_{a_i,b_i} - \vec{f}_{a_i,b'}) > 0$$

Therefore, a training data set is generated to feed the linear regression algorithm to learn the weights:

$$D = \{(x_i,y_i)\}_{i=1}^n \text{ where } x_i = (\vec{f}_{a_i,b_i} - \vec{f}_{a_i,b_{j\neq i}}) \quad y_i = 1$$

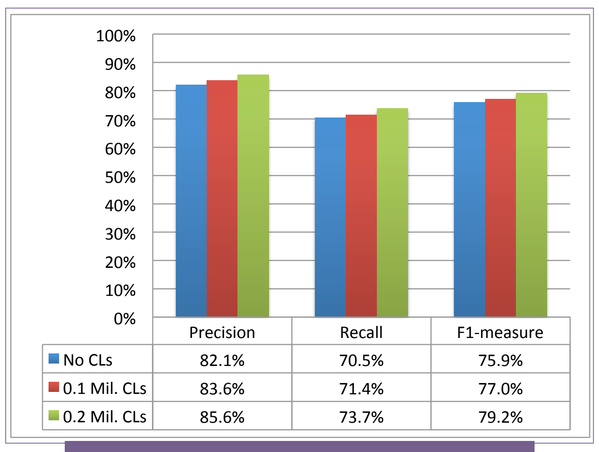
Experiments

The datasets of English Wikipedia (4 million articles) and Chinese Wikipedia (499 thousand articles) that are archived in August 2012 has been used to evaluate the proposed approach. There are 239,309 cross-lingual links between two wikis.

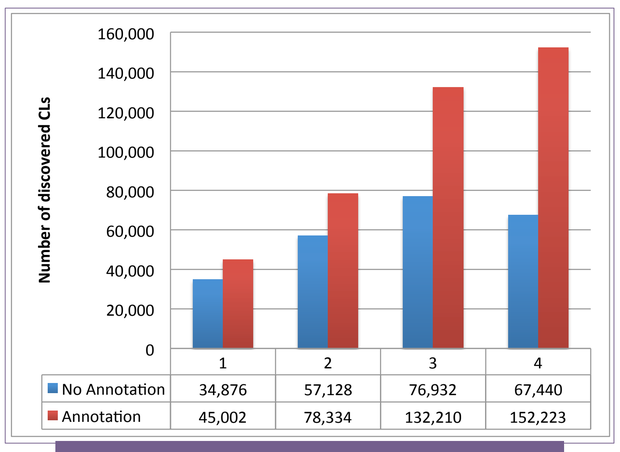
Results of cross-lingual link prediction

		Before Annotation			After Annotation		
#Seed CLs	Model	Precision	Reccall	F1-measure	Precision	Recall	F1-measure
0.05 Mil. CLs	SVM	92.1	35.0	50.7	78.5	37.2	50.5
	RM	93.3	36.0	**52.0**	92.4	38.6	**54.5**
0.10 Mil. CLs	SVM	79.7	35.0	48.6	86.9	50.4	63.8
	RM	84.6	37.4	**51.9**	96.6	49.3	**65.3**
0.15 Mil. CLs	SVM	80.9	35.9	49.7	88.1	57.3	69.5
	RM	93.5	38.2	**54.2**	93.7	56.2	**70.2**
0.20 Mil. CLs	SVM	84.7	37.3	51.8	88.8	68.1	77.1
	RM	94.5	37.9	**54.1**	95.9	67.2	**79.0**

Our regression model (RM) is compared with SVM classification model.



Results of concept annotation



Results of incremental linking