Heidelberg Institute for
Theoretical Studies

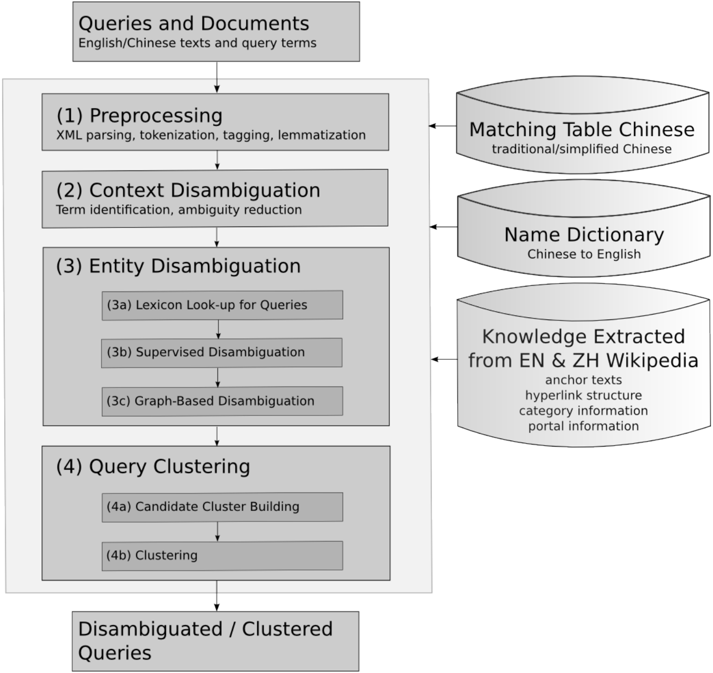
HITS

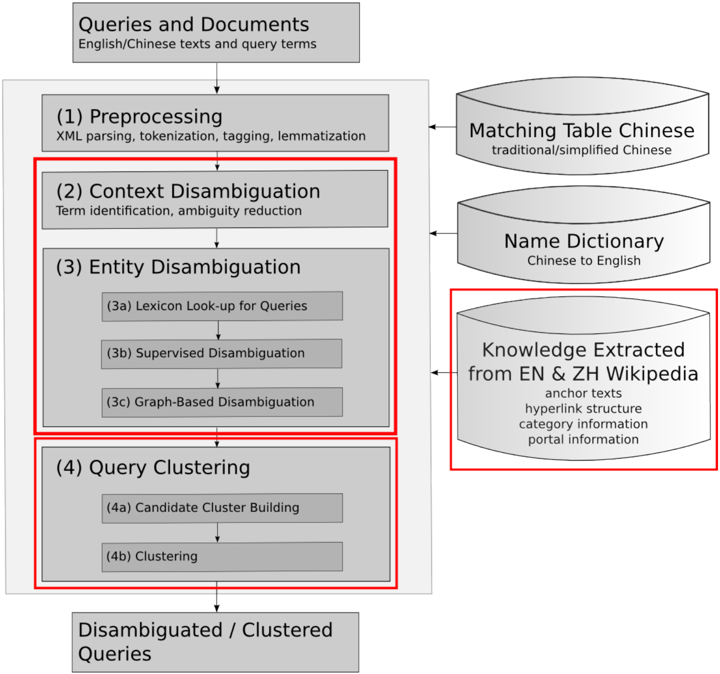
HITS’ Cross-lingual Entity Linking System at TAC 2011

One Model for All Languages

Angela Fahrni, Michael Strube, Vivi Nastase

Heidelberg Institute for Theoretical Studies gGmbH
Heidelberg, Germany


Queries and Documents
English/Chinese texts and query terms
(1) Preprocessing
XML parsing, tokenization, tagging, lemmatization
(2) Context Disambiguation
Term identification, ambiguity reduction
(3) Entity Disambiguation
(3a) Lexicon Look-up for Queries
(3b) Supervised Disambiguation
(3c) Graph-Based Disambiguation
(4) Query Clustering
(4a) Candidate Cluster Building
(4b) Clustering
Disambiguated / Clustered Queries
Matching Table Chinese
traditional/simplified Chinese
Name Dictionary
Chinese to English
Knowledge Extracted from EN & ZH Wikipedia
anchor texts
hyperlink structure
category information
portal information



Queries and Documents
English/Chinese texts and query terms

(1) Preprocessing
XML parsing, tokenization, tagging, lemmatization

(2) Context Disambiguation
Term identification, ambiguity reduction

(3) Entity Disambiguation
(3a) Lexicon Look-up for Queries
(3b) Supervised Disambiguation
(3c) Graph-Based Disambiguation

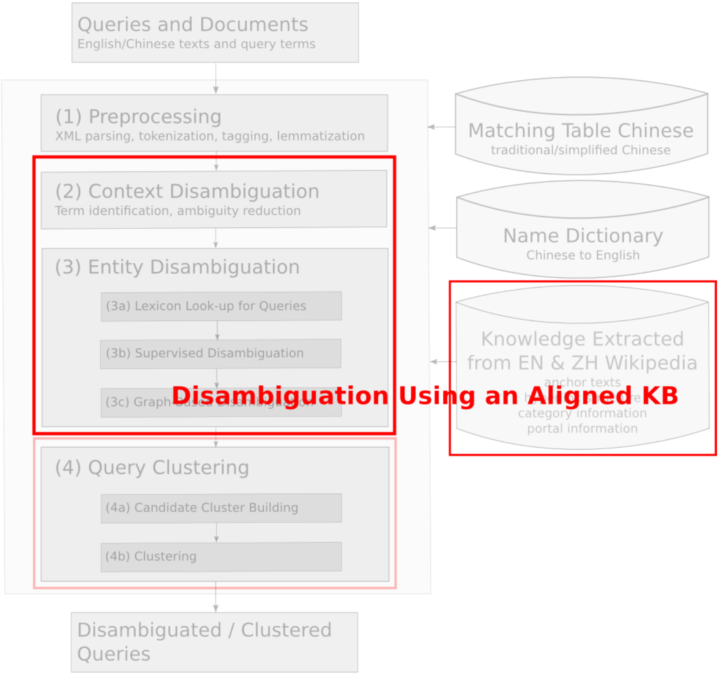
(4) Query Clustering
(4a) Candidate Cluster Building
(4b) Clustering

Disambiguated / Clustered Queries

Matching Table Chinese
traditional/simplified Chinese

Name Dictionary
Chinese to English

Knowledge Extracted from EN & ZH Wikipedia
anchor texts
hyperlink structure
category information
portal information



Disambiguation Using an Aligned KB

Queries and Documents
English/Chinese texts and query terms

(1) Preprocessing
XML parsing, tokenization, tagging, lemmatization

(2) Context Disambiguation
Term identification, ambiguity reduction

(3) Entity Disambiguation

(3a) Lexicon Look-up for Queries

(3b) Supervised Disambiguation

(3c) Graph-Based Disambiguation

(4) Query Clustering

(4a) Candidate Cluster Building

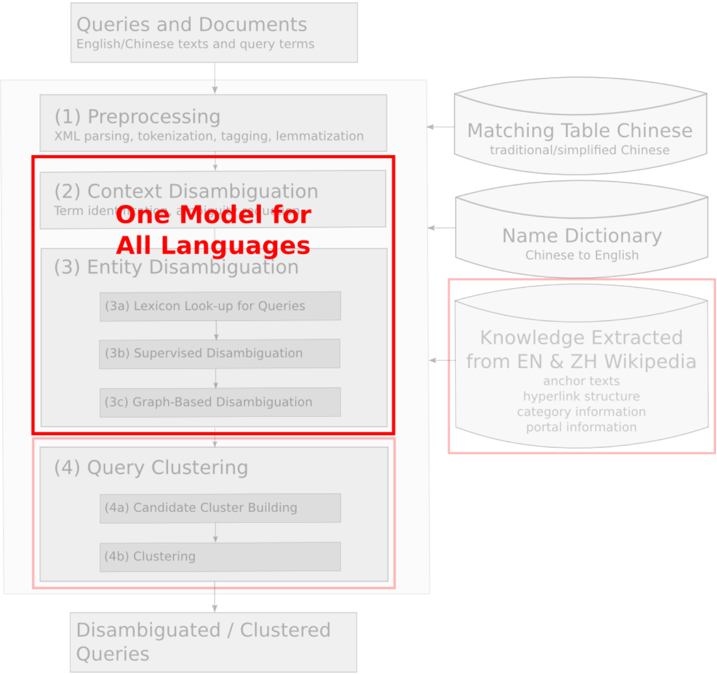
(4b) Clustering

Disambiguated / Clustered Queries

Matching Table Chinese
traditional/simplified Chinese

Name Dictionary
Chinese to English

Knowledge Extracted from EN & ZH Wikipedia
anchor texts
hyperlink structure
category information
portal information



Queries and Documents
English/Chinese texts and query terms

(1) Preprocessing
XML parsing, tokenization, tagging, lemmatization

(2) Context Disambiguation
Term identification, ambiguity reduction

(3) Entity Disambiguation

(3a) Lexicon Look-up for Queries

(3b) Supervised Disambiguation

(3c) Graph-Based Disambiguation

One Model for All Languages

(4) Query Clustering

(4a) Candidate Cluster Building

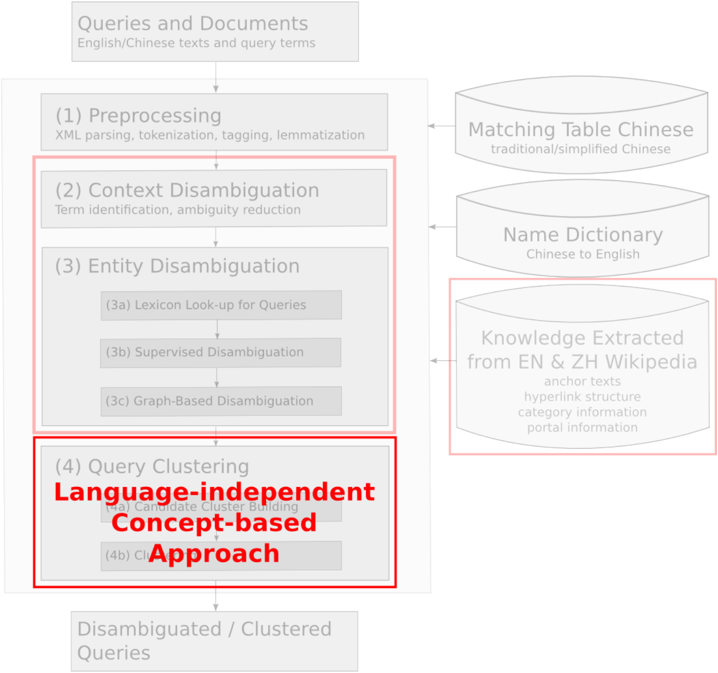
(4b) Clustering

Disambiguated / Clustered Queries

Matching Table Chinese
traditional/simplified Chinese

Name Dictionary
Chinese to English

Knowledge Extracted from EN & ZH Wikipedia
anchor texts
hyperlink structure
category information
portal information



Queries and Documents
English/Chinese texts and query terms

(1) Preprocessing
XML parsing, tokenization, tagging, lemmatization

(2) Context Disambiguation
Term identification, ambiguity reduction

(3) Entity Disambiguation

(3a) Lexicon Look-up for Queries

(3b) Supervised Disambiguation

(3c) Graph-Based Disambiguation

(4) Query Clustering

(4a) Candidate Cluster Building

(4b) Cl

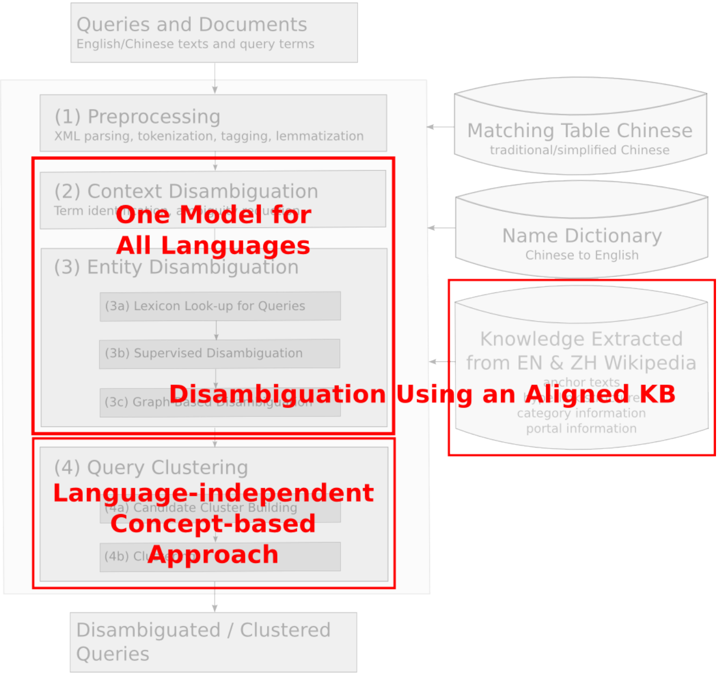
Language-independent Concept-based Approach

Matching Table Chinese
traditional/simplified Chinese

Name Dictionary
Chinese to English

Knowledge Extracted from EN & ZH Wikipedia
anchor texts
hyperlink structure
category information
portal information

Disambiguated / Clustered Queries



Queries and Documents
English/Chinese texts and query terms

(1) Preprocessing
XML parsing, tokenization, tagging, lemmatization

(2) Context Disambiguation
Term identification, ambiguity reduction

(3) Entity Disambiguation

(3a) Lexicon Look-up for Queries

(3b) Supervised Disambiguation

(3c) Graph-based Disambiguation

(4) Query Clustering

(4a) Candidate Cluster Building

(4b) Clustering

Disambiguated / Clustered Queries

Matching Table Chinese
traditional/simplified Chinese

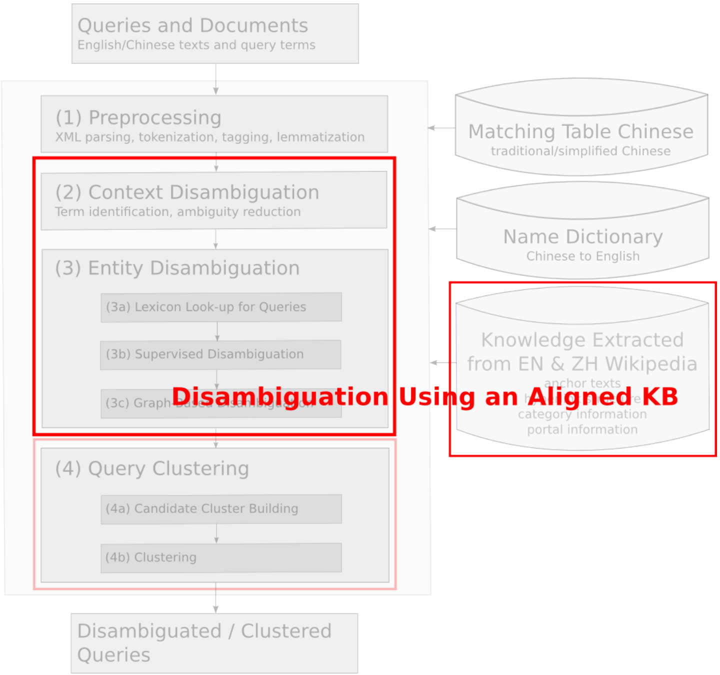
Name Dictionary
Chinese to English

Knowledge Extracted from EN & ZH Wikipedia
anchor texts
hyperlink structure
category information
portal information

One Model for All Languages

Disambiguation Using an Aligned KB

Language-independent Concept-based Approach



Queries and Documents
English/Chinese texts and query terms

(1) Preprocessing
XML parsing, tokenization, tagging, lemmatization

(2) Context Disambiguation
Term identification, ambiguity reduction

(3) Entity Disambiguation

(3a) Lexicon Look-up for Queries

(3b) Supervised Disambiguation

(3c) Graph-Based Disambiguation

(4) Query Clustering

(4a) Candidate Cluster Building

(4b) Clustering

Disambiguated / Clustered Queries

Matching Table Chinese
traditional/simplified Chinese

Name Dictionary
Chinese to English

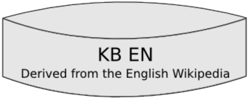
Knowledge Extracted from EN & ZH Wikipedia
anchor texts
hyperlink structure
category information
portal information

Disambiguation Using an Aligned KB

Strategies for Cross-lingual Entity Disambiguation

Query term: 艾尔沙德

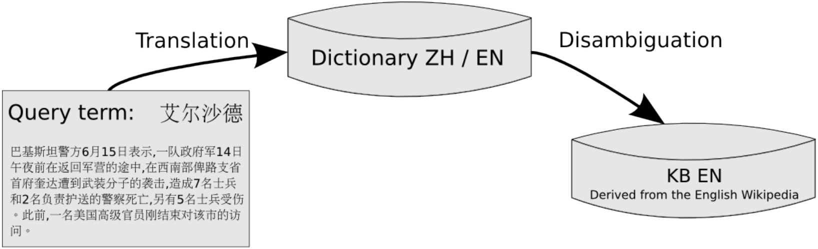
巴基斯坦警方6月15日表示,一队政府军14日午夜前在返回军营的途中,在西南部俾路支省首府奎达遭到武装分子的袭击,造成7名士兵和2名负责护送的警察死亡,另有5名士兵受伤。此前,一名美国高级官员刚结束对该市的访问。



KB EN

Derived from the English Wikipedia

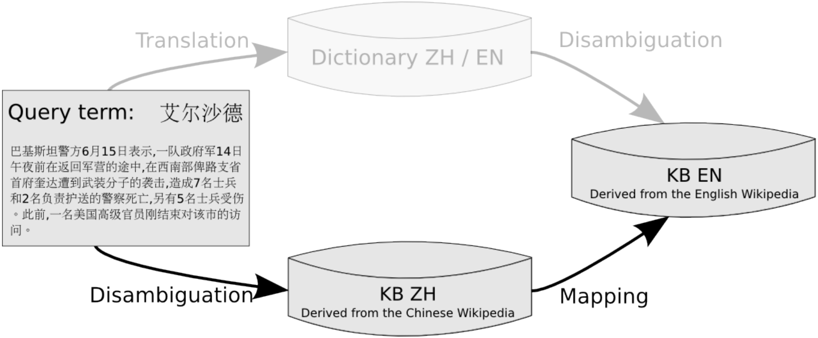
Strategies for Cross-lingual Entity Disambiguation



Query term: 艾尔沙德

巴基斯坦警方6月15日表示,一队政府军14日午夜前在返回军营的途中,在西南部俾路支省首府奎达遭到武装分子的袭击,造成7名士兵和2名负责护送的警察死亡,另有5名士兵受伤。此前,一名美国高级官员刚结束对该市的访问。

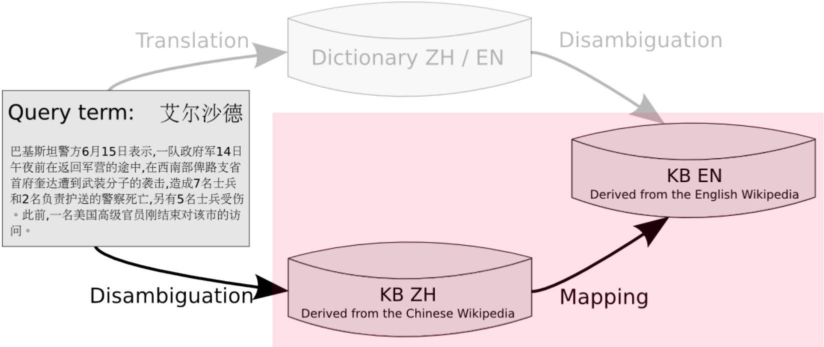
Strategies for Cross-lingual Entity Disambiguation



巴基斯坦警方6月15日表示,一队政府军14日午夜前在返回军营的途中,在西南部俾路支省首府奎达遭到武装分子的袭击,造成7名士兵和2名负责护送的警察死亡,另有5名士兵受伤。此前,一名美国高级官员刚结束对该市的访问。

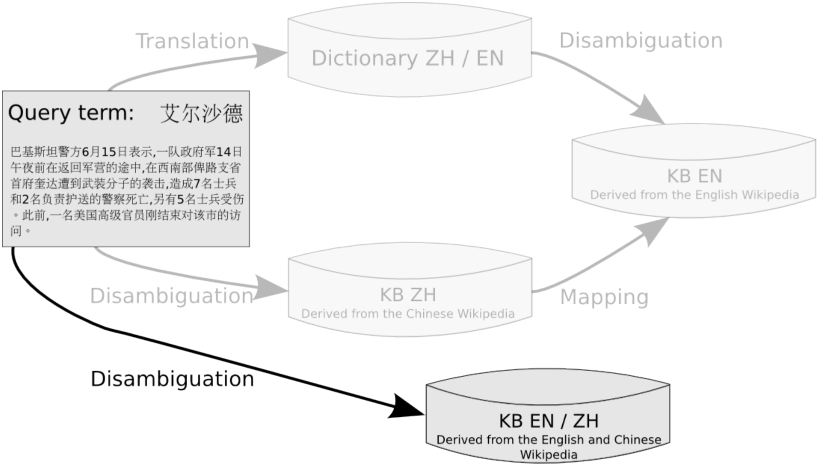
KB EN
Derived from the English Wikipedia
Disambiguation
KB ZH
Derived from the Chinese Wikipedia
Mapping

Strategies for Cross-lingual Entity Disambiguation


Query term: 艾尔沙德

巴基斯坦警方6月15日表示,一队政府军14日午夜前在返回军营的途中,在西南部俾路支省首府奎达遭到武装分子的袭击,造成7名士兵和2名负责护送的警察死亡,另有5名士兵受伤。此前,一名美国高级官员刚结束对该市的访问。

Strategies for Cross-lingual Entity Disambiguation



Query term: 艾尔沙德

巴基斯坦警方6月15日表示,一队政府军14日午夜前在返回军营的途中,在西南部俾路支省首府奎达遭到武装分子的袭击,造成7名士兵和2名负责护送的警察死亡,另有5名士兵受伤。此前,一名美国高级官员刚结束对该市的访问。

Translation

Dictionary ZH / EN

Disambiguation

KB EN

Derived from the English Wikipedia

Disambiguation

KB ZH

Derived from the Chinese Wikipedia

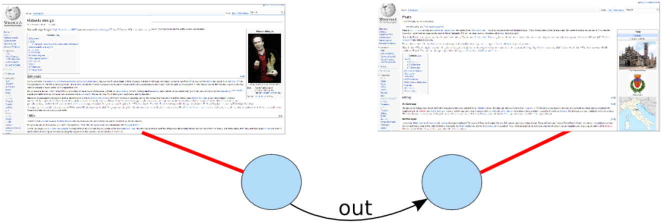
Mapping

Disambiguation

KB EN / ZH

Derived from the English and Chinese Wikipedia

Knowledge base



Token level information
Associated tokens

Category Information
Associated categories

Portal Information
Associated portals

Lexical Realizations
Article names
Redirects
Disambiguation pages
Bold terms

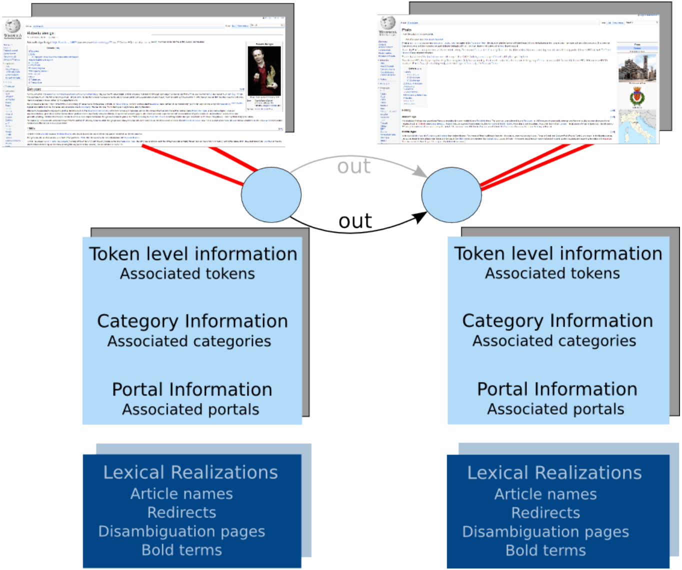
Token level information
Associated tokens

Category Information
Associated categories

Portal Information
Associated portals

Lexical Realizations
Article names
Redirects
Disambiguation pages
Bold terms

Knowledge base



Token level information
Associated tokens

Category Information
Associated categories

Portal Information
Associated portals

Lexical Realizations
Article names
Redirects
Disambiguation pages
Bold terms

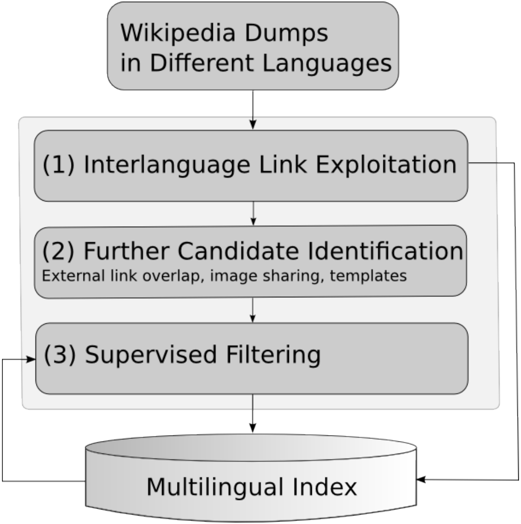
Token level information
Associated tokens

Category Information
Associated categories

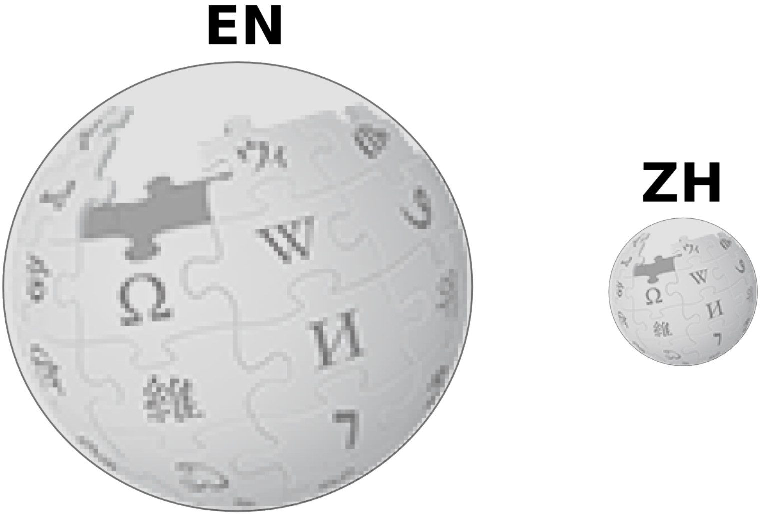
Portal Information
Associated portals

Lexical Realizations
Article names
Redirects
Disambiguation pages
Bold terms

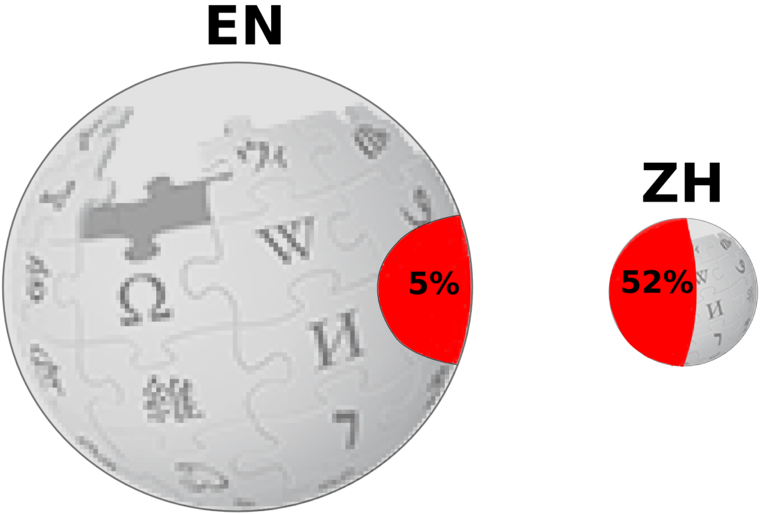
Mapping between the English and Chinese Wikipedia Versions


Wikipedia Dumps
in Different Languages
(1) Interlanguage Link Exploitation
(2) Further Candidate Identification
External link overlap, image sharing, templates
(3) Supervised Filtering
Multilingual Index

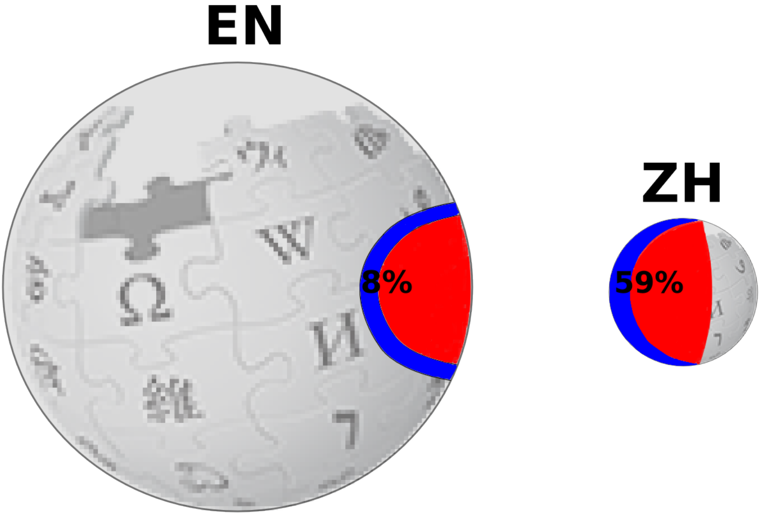
Mapping between the English and Chinese Wikipedia Versions



Mapping between the English and Chinese Wikipedia Versions

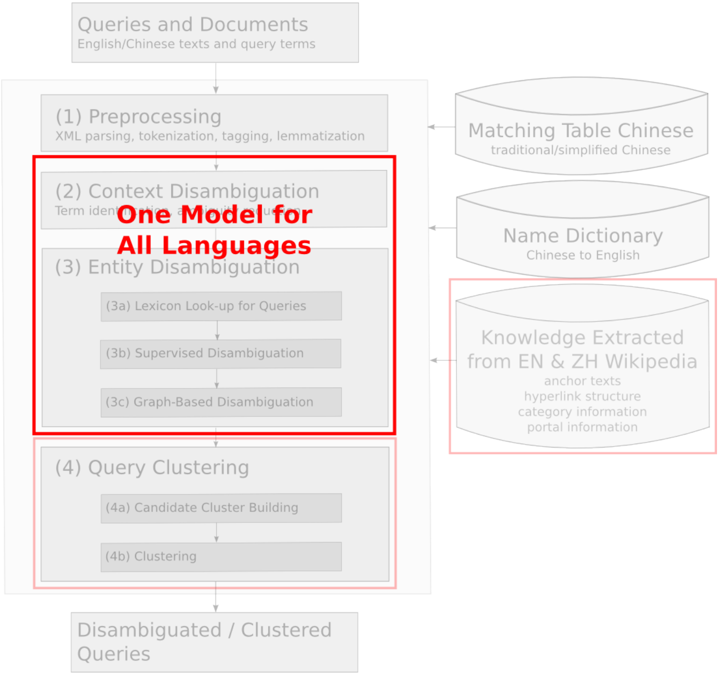


Mapping between the English and Chinese Wikipedia Versions



Coverage after Lexicon Lookup

	Training Set	
	Coverage	Average Ambiguity
EN	0.928	16.92
ZH	0.867	5.56
ZH_Lex	0.930	22.86



Queries and Documents
English/Chinese texts and query terms

(1) Preprocessing
XML parsing, tokenization, tagging, lemmatization

(2) Context Disambiguation
Term identification, ambiguity reduction

One Model for All Languages

(3) Entity Disambiguation

(3a) Lexicon Look-up for Queries

(3b) Supervised Disambiguation

(3c) Graph-Based Disambiguation

(4) Query Clustering

(4a) Candidate Cluster Building

(4b) Clustering

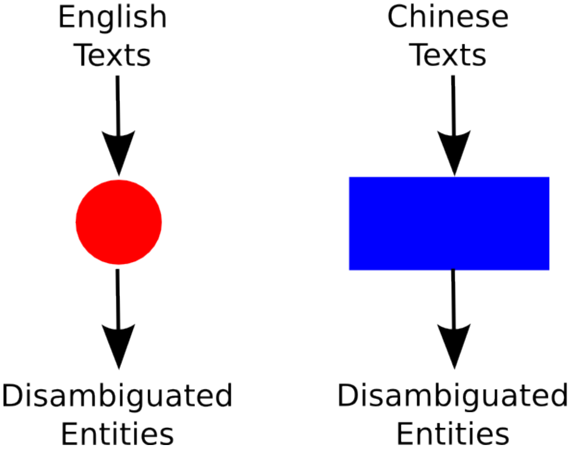
Disambiguated / Clustered Queries

Matching Table Chinese
traditional/simplified Chinese

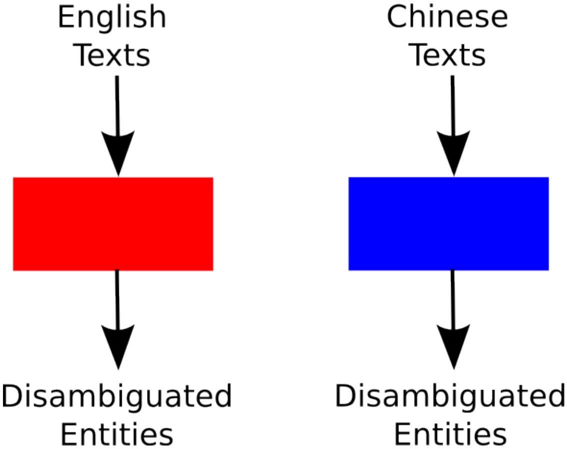
Name Dictionary
Chinese to English

Knowledge Extracted from EN & ZH Wikipedia
anchor texts
hyperlink structure
category information
portal information

Different Techniques for Different Languages



Different Models for Different Languages



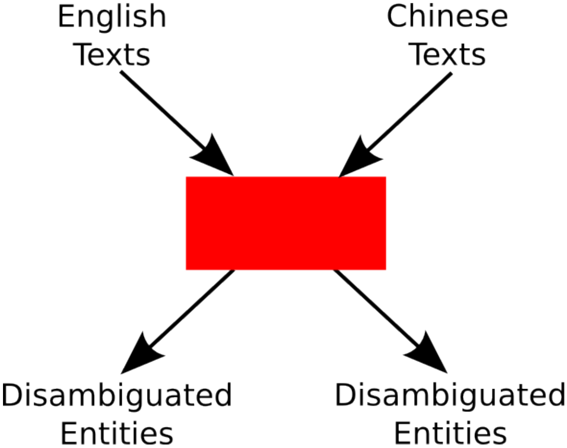
English Texts

Chinese Texts

Disambiguated Entities

Disambiguated Entities

One Model for All Languages



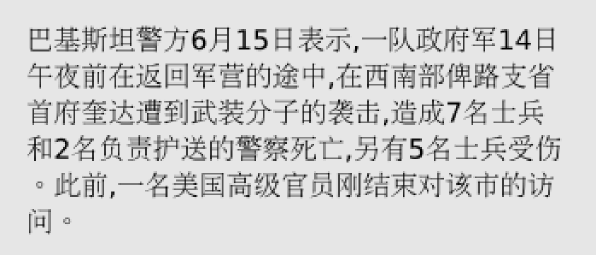
English Texts

Chinese Texts

Disambiguated Entities

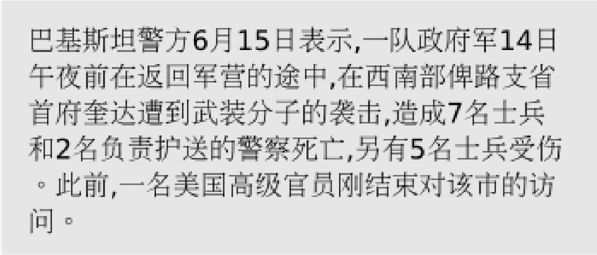
Disambiguated Entities

Context Disambiguation Approach

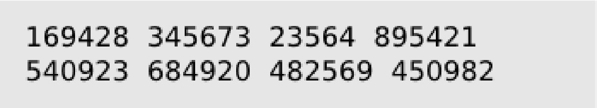


巴基斯坦警方6月15日表示,一队政府军14日午夜前在返回军营的途中,在西南部俾路支省首府奎达遭到武装分子的袭击,造成7名士兵和2名负责护送的警察死亡,另有5名士兵受伤。此前,一名美国高级官员刚结束对该市的访问。

Context Disambiguation Approach

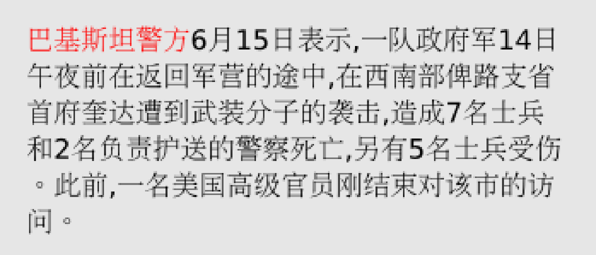


巴基斯坦警方6月15日表示,一队政府军14日
午夜前在返回军营的途中,在西南部俾路支省
首府奎达遭到武装分子的袭击,造成7名士兵
和2名负责护送的警察死亡,另有5名士兵受伤
。此前,一名美国高级官员刚结束对该市的访
问。



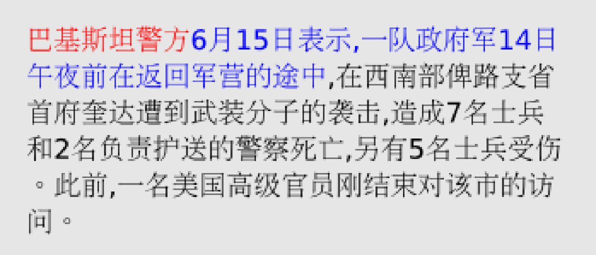
169428 345673 23564 895421
540923 684920 482569 450982

Context Disambiguation Approach



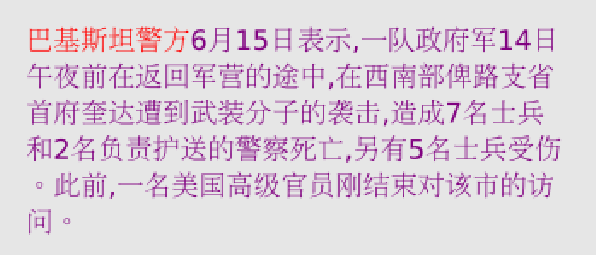
巴基斯坦警方6月15日表示,一队政府军14日午夜前在返回军营的途中,在西南部俾路支省首府奎达遭到武装分子的袭击,造成7名士兵和2名负责护送的警察死亡,另有5名士兵受伤。此前,一名美国高级官员刚结束对该市的访问。

Context Disambiguation Approach



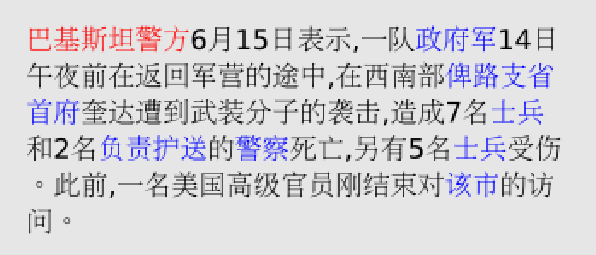
巴基斯坦警方6月15日表示,一队政府军14日午夜前在返回军营的途中,在西南部俾路支省首府奎达遭到武装分子的袭击,造成7名士兵和2名负责护送的警察死亡,另有5名士兵受伤。此前,一名美国高级官员刚结束对该市的访问。

Context Disambiguation Approach



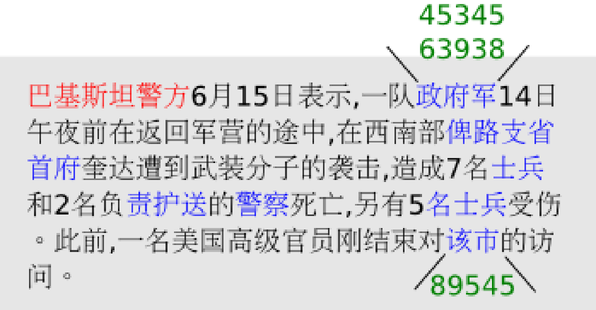
巴基斯坦警方6月15日表示,一队政府军14日午夜前在返回军营的途中,在西南部俾路支省首府奎达遭到武装分子的袭击,造成7名士兵和2名负责护送的警察死亡,另有5名士兵受伤。此前,一名美国高级官员刚结束对该市的访问。

Context Disambiguation Approach



巴基斯坦警方6月15日表示,一队政府军14日午夜前在返回军营的途中,在西南部俾路支省首府奎达遭到武装分子的袭击,造成7名士兵和2名负责护送的警察死亡,另有5名士兵受伤。此前,一名美国高级官员刚结束对该市的访问。

Context Disambiguation Approach



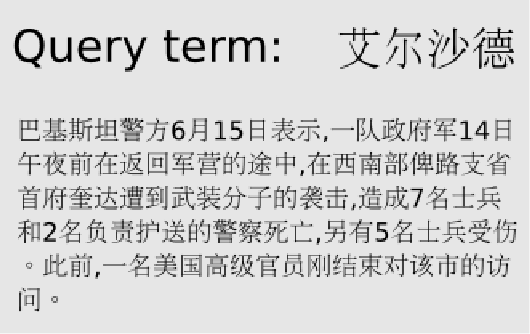
巴基斯坦警方6月15日表示,一队政府军14日午夜前在返回军营的途中,在西南部俾路支省首府奎达遭到武装分子的袭击,造成7名士兵和2名负责护送的警察死亡,另有5名士兵受伤。此前,一名美国高级官员刚结束对该市的访问。

45345
63938
89545

Context Disambiguation Approach

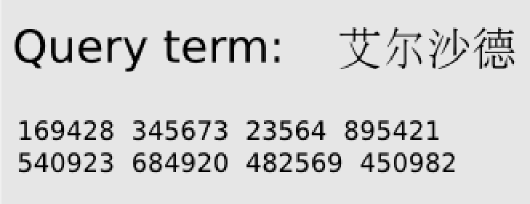
- Training of a supervised model (SVM) using English training instances
- Training instances are derived from the internal hyperlinks in Wikipedia
  - Positive instances: linked Wikipedia articles
  - Negative instances: other Wikipedia articles the terms can refer to
- Model is used for English and Chinese

Entity Disambiguation



Query term: 艾尔沙德

巴基斯坦警方6月15日表示,一队政府军14日
午夜前在返回军营的途中,在西南部俾路支省
首府奎达遭到武装分子的袭击,造成7名士兵
和2名负责护送的警察死亡,另有5名士兵受伤
。此前,一名美国高级官员刚结束对该市的访
问。



Query term: 艾尔沙德

169428 345673 23564 895421
540923 684920 482569 450982

Entity Disambiguation Approach

- Training instances are derived from the training data provided by TAC
  - Positive instances: correct entity for the query terms
  - Negative instances: other candidate entities for the query terms
- Training of a supervised model (SVM) based on
  - English and Chinese training instances
  - Chinese training instances
  - English instances
- Application of the models to both languages

One-Model-for-All-Languages: Features

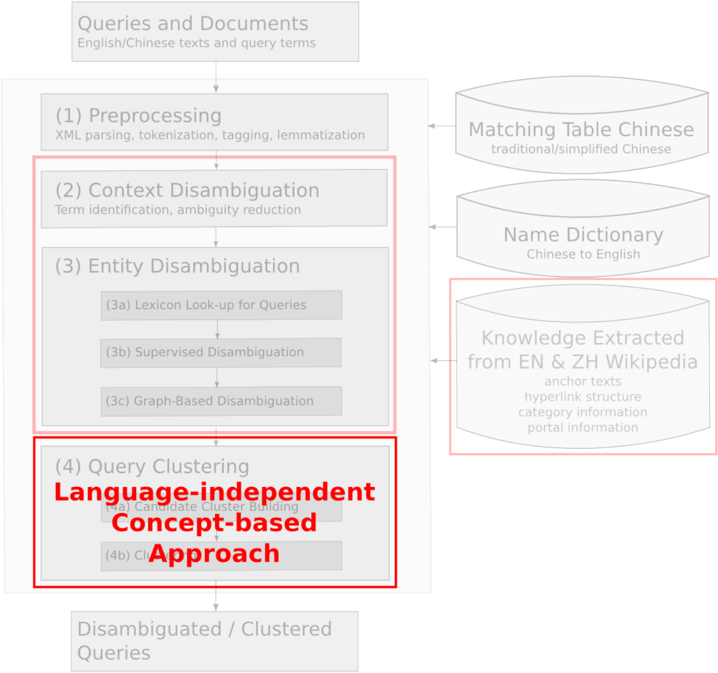
- Abstraction from particular languages, i.e. no lexical features
- Relatedness measures
- Concept-based features

One-Model-for-All-Languages: Features

- Prior probability
- Distance between term and name of the respective Wikipedia article
- Context fit on conceptual level
  - Vector-based features based on concepts, categories, lists, portals
  - Average, maximum and minimum derived from pairwise calculated relatedness measures
    - Relatedness measures based on outgoing links, incoming links, categories
- Context fit on token level: Cosine similarity based on nouns, verbs, adjectives between:
  - Whole text and respective Wikipedia article
  - Surrounding context and local context of hyperlinks pointing to the respective page in Wikipedia

Results Micro-Average
Different Training Approaches

	Micro-Average
HITS_Trained_EN_ZH	0.785
HITS_Trained_SEP	0.789
HITS_Trained_ZH	0.786



Queries and Documents

English/Chinese texts and query terms

(1) Preprocessing

XML parsing, tokenization, tagging, lemmatization

(2) Context Disambiguation

Term identification, ambiguity reduction

(3) Entity Disambiguation

(3a) Lexicon Look-up for Queries

(3b) Supervised Disambiguation

(3c) Graph-Based Disambiguation

(4) Query Clustering

(4a) Candidate Cluster Building

(4b) Cl

Language-independent Concept-based Approach

Disambiguated / Clustered Queries

Matching Table Chinese

traditional/simplified Chinese

Name Dictionary

Chinese to English

Knowledge Extracted from EN & ZH Wikipedia

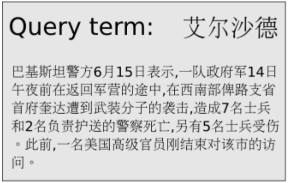
anchor texts

hyperlink structure

category information

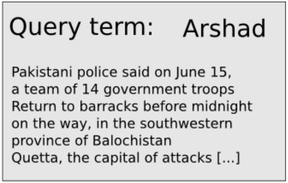
portal information

Language-independent Concept-based Representation



Query term: 艾尔沙德

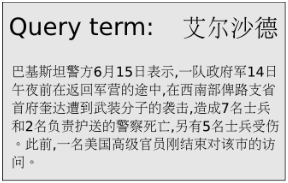
巴基斯坦警方6月15日表示,一队政府军14日
午夜前在返回军营的途中,在西南部俾路支省
首府奎达遭到武装分子的袭击,造成7名士兵
和2名负责护送的警察死亡,另有5名士兵受伤
。此前,一名美国高级官员刚结束对该市的访
问。



Query term: Arshad

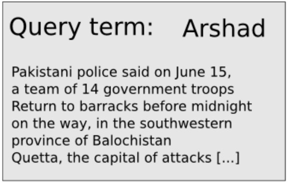
Pakistani police said on June 15,
a team of 14 government troops
Return to barracks before midnight
on the way, in the southwestern
province of Balochistan
Quetta, the capital of attacks [...]

Language-independent Concept-based Representation



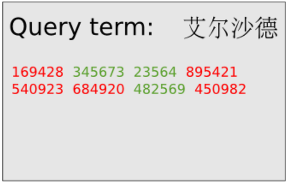
Query term: 艾尔沙德

巴基斯坦警方6月15日表示,一队政府军14日午夜前在返回军营的途中,在西南部俾路支省首府奎达遭到武装分子的袭击,造成7名士兵和2名负责护送的警察死亡,另有5名士兵受伤。此前,一名美国高级官员刚结束对该市的访问。



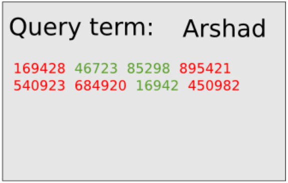
Query term: Arshad

Pakistani police said on June 15,
a team of 14 government troops
Return to barracks before midnight
on the way, in the southwestern
province of Balochistan
Quetta, the capital of attacks [...]



Query term: 艾尔沙德

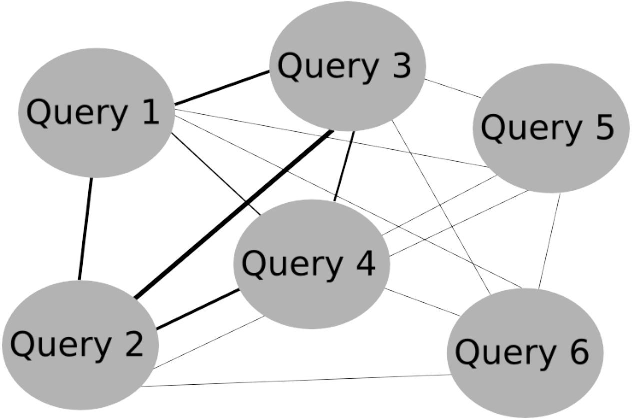
169428 345673 23564 895421
540923 684920 482569 450982



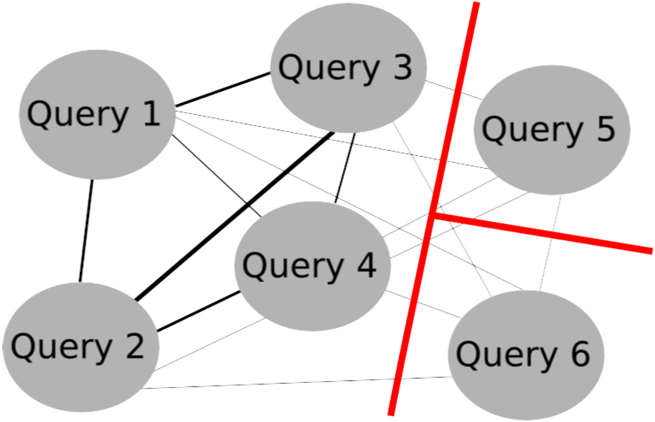
Query term: Arshad

169428 46723 85298 895421
540923 684920 16942 450982

Graph



Clustering



Learning the Edge Weights

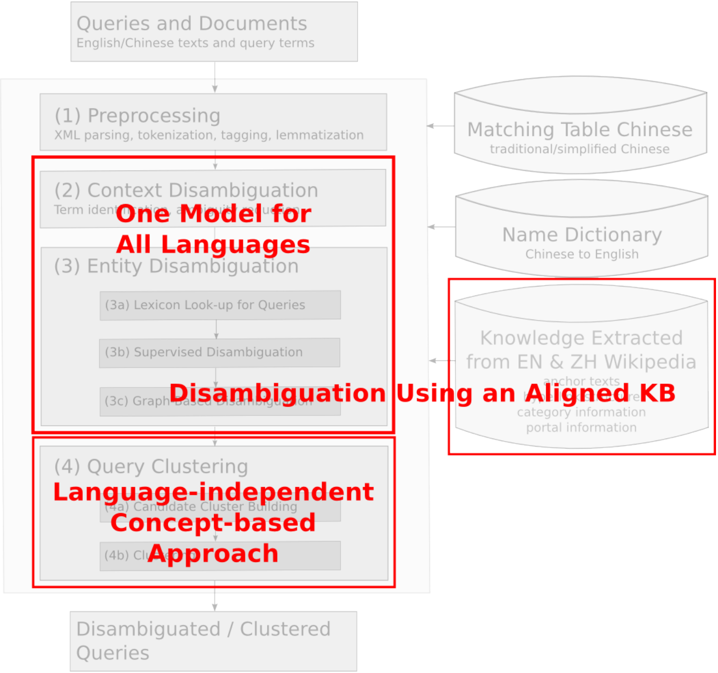
- Binary classifier (SMV) with the classes:
  - In same cluster
  - Not in the same cluster
- Confidence values as edge weights
- Features: Cosine similarities between local contexts and whole texts based on:
  - Identified concepts
  - Identified concepts extended by incoming and outgoing links
  - Categories associated with the concepts
  - Lists associated with the concepts
  - Portals associated with the concepts

Shortcomings of Graph-based Clustering Approach

- Construction of the graph is inefficient
  - Pairwise comparisons using different similarity measures
- Few data to optimize the parameters
- Temporary solution: String match heuristic

Results

	Micro-Average	**Precision (B^3)**	**Recall (B^3)**	**F1 (B^3)**
Best System				0.788
Median				0.675
HITS	**0.785**	**0.700**	**0.763**	**0.730**



Queries and Documents
English/Chinese texts and query terms

(1) Preprocessing
XML parsing, tokenization, tagging, lemmatization

(2) Context Disambiguation
Term identification, ambiguity reduction

One Model for All Languages

(3) Entity Disambiguation

(3a) Lexicon Look-up for Queries

(3b) Supervised Disambiguation

(3c) Graph-based Disambiguation

(4) Query Clustering

Language-independent Concept-based Approach

(4a) Candidate Cluster Building

(4b) Clustering

Disambiguated / Clustered Queries

Matching Table Chinese
traditional/simplified Chinese

Name Dictionary
Chinese to English

Knowledge Extracted from EN & ZH Wikipedia
anchor texts
hyperlink structure
category information
portal information

Disambiguation Using an Aligned KB

Future Work

- Experiments with other languages
  - We participated in the cross-lingual link discovery task organized by NTCIR: good results for all subtasks (EN-to-Chinese, EN-to-Korean, EN-to-Japanese)

- Efficient way to cluster terms: Circumventing pairwise comparisons

Thank you!

Heidelberger Institut für Theoretische Studien HITS

COSYNE

SEVENTH FRAMEWORK PROGRAMME

Angela Fahrni
angela.fahrni@h-its.org
http://www.h-its.org