西南交通大学 Southwest Jiaotong University

Research on Entity Linking of Chinese Micro-Blog

Zhen Jia

Contents

LOGO

1. Key Problems

- **(1) Entity name may be abbreviation or alias**

  E.g.1 “小皇帝在今天的比赛中发威了”(*King James did well in today’s game*)

  King James is the alias of James Brown.

- **(2) There are many noises or errors in Micro-blog**

  E.g.2 “周杰轮”→ “周杰伦”

- **(3) Properties or attributes in knowledge base are not uniform**

  E.g.3 *Some properties are in English, some are in Chinese*.

- **(4) Different entities have the same name**

  E.g.4 *黑龙江(Heilongjiang) may refer to a river or a province.*

  E.g.5 *The person name “王维” (**Wang Wei**) in Micro-blog sometimes does not refer to a poet “王维” and many people can use the same name.*

Contents

2. Methods

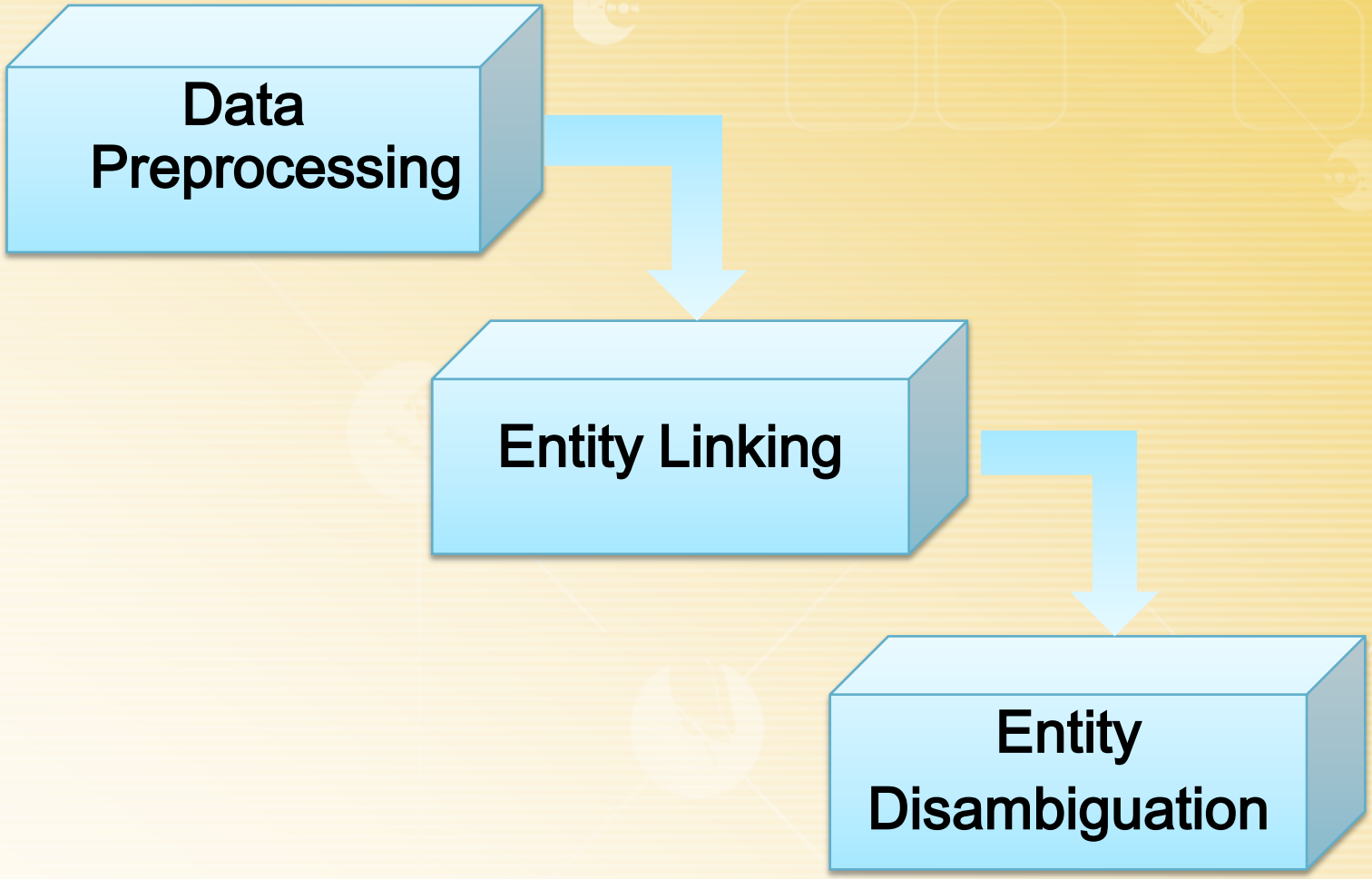
(1) Construct synonymy thesaurus to solve problem that one entity has different names.

(2) Reconstruct knowledge base and remove noises and errors in it.

(3) Use improved Pinyin edit distance algorithm to distinguish wrongly written characters .

(4) Compute linking value to disambiguate an entity name which refers to several different things in knowledge base.

Pipeline of method



Step 1.Data Preprocessing

Reconstruct Knowledge base

- Use sample data of NLP&CC2013 to label the categories of entities.
- Select high frequency attributes as representative attributes of categories.

Construct Synonymy Thesaurus

- Extract synonymous entries from Baidubaike to construct synonymy thesaurus.

Get Popularity of Entities

- Extract visit times of entries from Baidubaike as the popularity of entities.

Preprocess Micro-blog Texts

- Use SWJTU Chinese segmentation system to preprocess Micro-blog text.

Step 2.Entity Linking

Input：an entity name in Micro-blog

Search entity names or their synonym directly in knowledge base

Use improved Pinyin edit distance algorithm to find entity names

Use literal matching algorithm to find entity names

Output：KB numbers of the entity name

Step 2.Entity Linking

Improved Pinyin Edit Distance

- Spell is set of consonants and vowels with similar pronunciation

Spell={(l,n),(l,r),(z,zh),(c,ch),(an,ang),(en,eng),(in,ing),(ang,ong), (si,ci)}

- Difference degree

$$\mathrm{DifferenceDegree}(I, I') = \begin{cases} 0.5, & I \in \mathrm{Spell}_i, I' \in \mathrm{Spell}_i \\ 1, & \mathrm{other} \end{cases}$$

- Difference degree of entity names

$$\mathrm{AllDifferenceDegree}(E_a, E_b) = \sum_{i=1}^{\max(m,n)} \mathrm{DifferenceDegree}_i$$

m is length of Pinyin of entity Ea. n is length of Pinyin of entity Eb.
If AllDifferenceDegree (Ea, Eb) ≥ 1, Ea and Eb are different entities.
If AllDifferenceDegree of (Ea, Eb) < 1, Ea and Eb are the same entities.

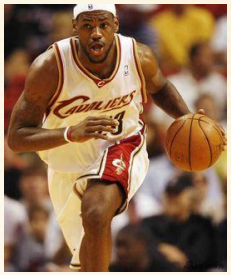
Step 3.Entity Disambiguation

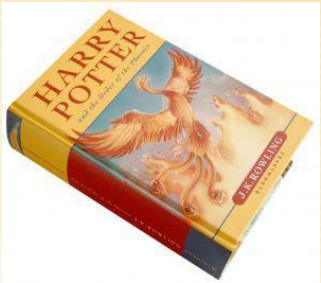
- Because Micro-blog is open to everyone and the content is very short, entity names in Micro-blog have the following characteristics:

Diversity

An entity may use many kinds of names such as full name, alias, or abbreviation.

Lebron, James, King, LBJ,…





Harry Porter

Ambiguity

A name may refer to different entities in different context of Micro-blog.

Step 3.Entity Disambiguation

Similarity

- Compute similarity between category of Micro-blog and candidate categories of entity in knowledge base.

Popularity

- Extract visit times from Baidubaike as popularity of the entity.

Linking Value

- Compute linking value of entity to decide which category the entity belongs to.

Step 3.Entity Disambiguation

Similarity

- Use Micro-blogs context and even all Micro-blogs in one topic as corpus, and count the words which are same as category words, such as 人物/体育/篮球(people/sports/basketball)、人物/体育/足球(people/sports/soccer)、人物/科学家/物理学家(people/scientist/physicist).
- Assign weight value λ (λ =1.0, 0.8, 0.7, …) to each category according to the number of words in each category.
- Compute the similarity $R(C_{xi}, C_y)(y = 1,2,\ldots)$ between the candidate category C_{xi} which entity Xi in knowledge base belongs to and the category C_y which entity in Micro-blog might belong to.
- Choose the maximum similarity value $\varphi_{xi} = \max\{R(C_{xi}, C_y), y = 1,2,\ldots\}$ as similarity result.

Step 3.Entity Disambiguation

Popularity

- Extract visit times from Baidubaike as popularity of entity.
- Assign a weight value β (β=1, 0.8, 0.7,..) to each popularity.

Linking Value

- Compute linking value *S(xi).*
- If *S(xi)* is maximum and *xi* is the entity linking to knowledge base.

$$S(x_i) = 0.6 \times \lambda \times \varphi_{xi} + 0.4 \times \beta$$

Contents

LOGO

3.Evaluation Result

Training Data

- Provided by NLPCC2014 and containing 177 Chinese Micro-blogs
- Extracted from Sina and containing 1000 microblogs

Testing Data

- Provided by NLPCC2014 and containing 1152 entity names.

System ID	Overall		in-KB			NIL		
	Correct #	Accuracy	Precision	Recall	F1	Precision	Recall	F1
4	527	0.8682	0.8078	0.8598	0.8330	0.9202	0.8746	0.8969

Contents

LOGO

4.Conclusion

NLP Preprocessing

- SWJTU Chinese segmentation system

Major Approaches

- Knowledge base reconstruction
- Synonymy thesaurus construction
- Improved Pinyin edit distance algorithm
- Literal matching algorithm
- Linking value computation based on similarity and popularity

Pros & Cros

- The method of entity linking is effective.
- The performance of entity disambiguation needs to be improved.

Appreciate your attention!
Any Questions?