# Construction of a Chinese Entity Linking Corpus

Reporter:Jiagen Shu
Jiagen Shu,Haotian Hui,Longhua Qian,Qiaoming Zhu
NLP Lab , Soochow University
http://nlp.suda.edu.cn

# Main Content

- Problem Description
- Related Work
- Annotation of the Corpus
- Baseline
- Future and Prospect

# Problem Description

Entity Linking:

Entity Linking task is a sub task of KBP (knowledge base population) task which was released by TAC (text analysis conference). It can help the task of slot filling.

Task definition:

Linking one entity mention to its corresponding entity in the knowledge base.

Definition of entity mention:

An entity mention is a reference to an entity. It can be some strings. It contains three types:Name Mention,

Nominal Mention, Pronoun Mentions

# Problem Description

Example:

[The vice chairman of Olympic committee] [Jingmin Liu] showed that [he] reached his target in this trip when he was interviewed by journalists.

[The vice chairman of Olympic committee] is a nominal mention.

[Jingmin Liu] is a name mention.

[he] is a pronoun mentions.

If this entity mention( [Jingmin Liu] ) has its corresponding entity in knowledge base, return the entity id, or return a NIL.

# Related Work

English entity linking corpus:
example:TAC2010

Cross-language entity linking corpus:
example:TAC2011

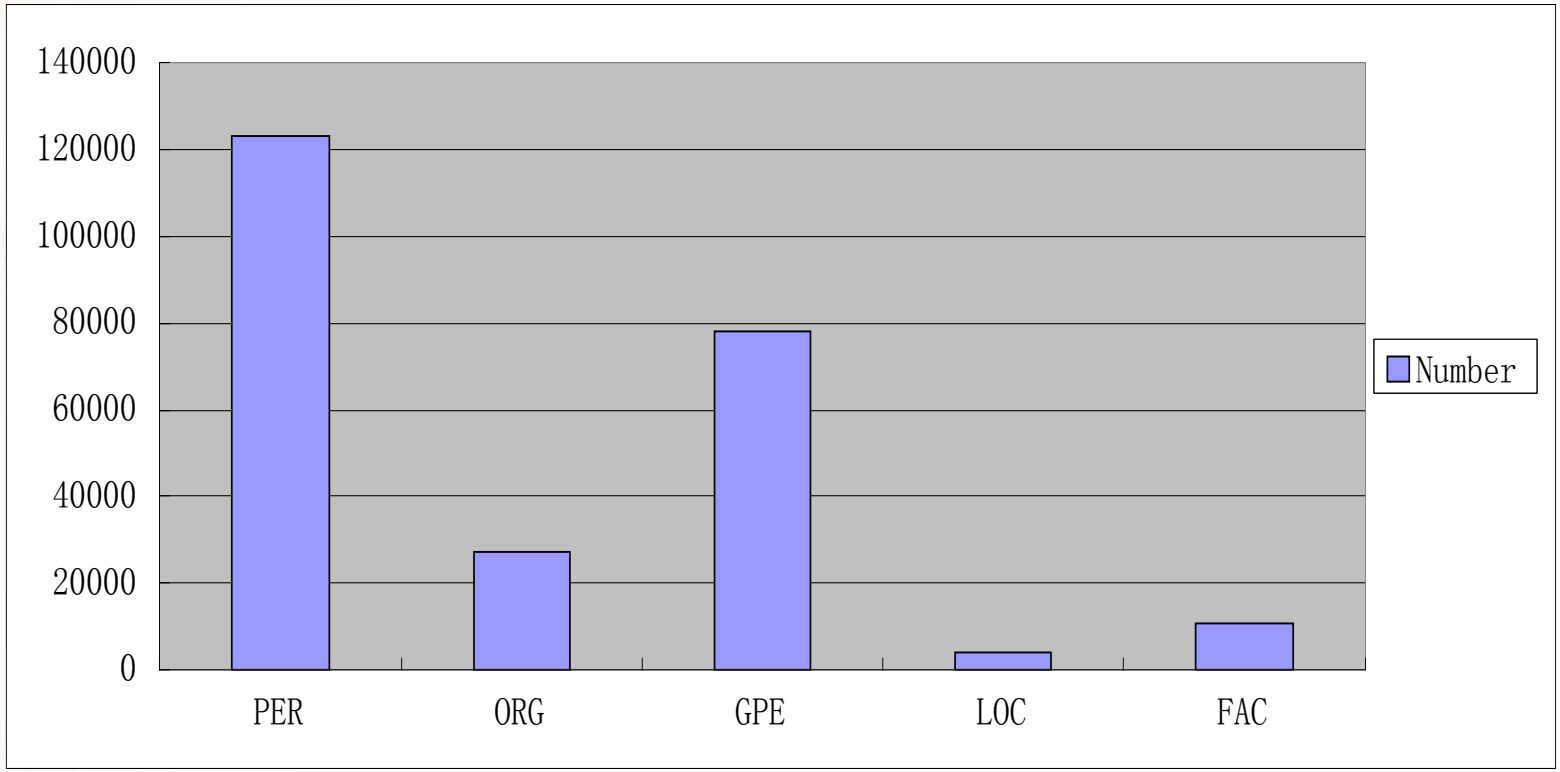
Chinese entity linking corpus:
example:NLPCC2013

Similar annotation:
example:Cucerzan(2007)　Csomai & Mihalcea（2008）
Milne & Witten（2008）Kulkarni(2009) Bentivogli(2010)

# Annotation of the corpus

1.The construction of the Chinese knowledge base.

Our knowledge base derived from Chinese Wikipedia. It is made of five kinds of entity type:PER(Person) ORG(Organization) GPE(Geo-Political Entities) FAC(Facility) LOC(Location).



# Annotation of the corpus

One example of entity in our knowledge base:

<entity wiki_title="王审知" type="PER" id="E205886" name="王审知">
<facts class="Emperorcn box">
<fact name="姓名">王审知</fact>
……
</facts>
<wiki_text><![CDATA[王审知
闽 太祖 王审知， 字信通， 一字详卿。光州 固始（今 河南 固始 ）
人……
]]></wiki_text>
</entity>

# Annotation of the corpus

2.Chinese corpus of ACE2005

It consists of 633 different texts from different domains.It contains 6,771 different entities.

3.The method of annotation

1). Automatic annotation

We think the entity whose string exactly matches the entity mention is the corresponding one.

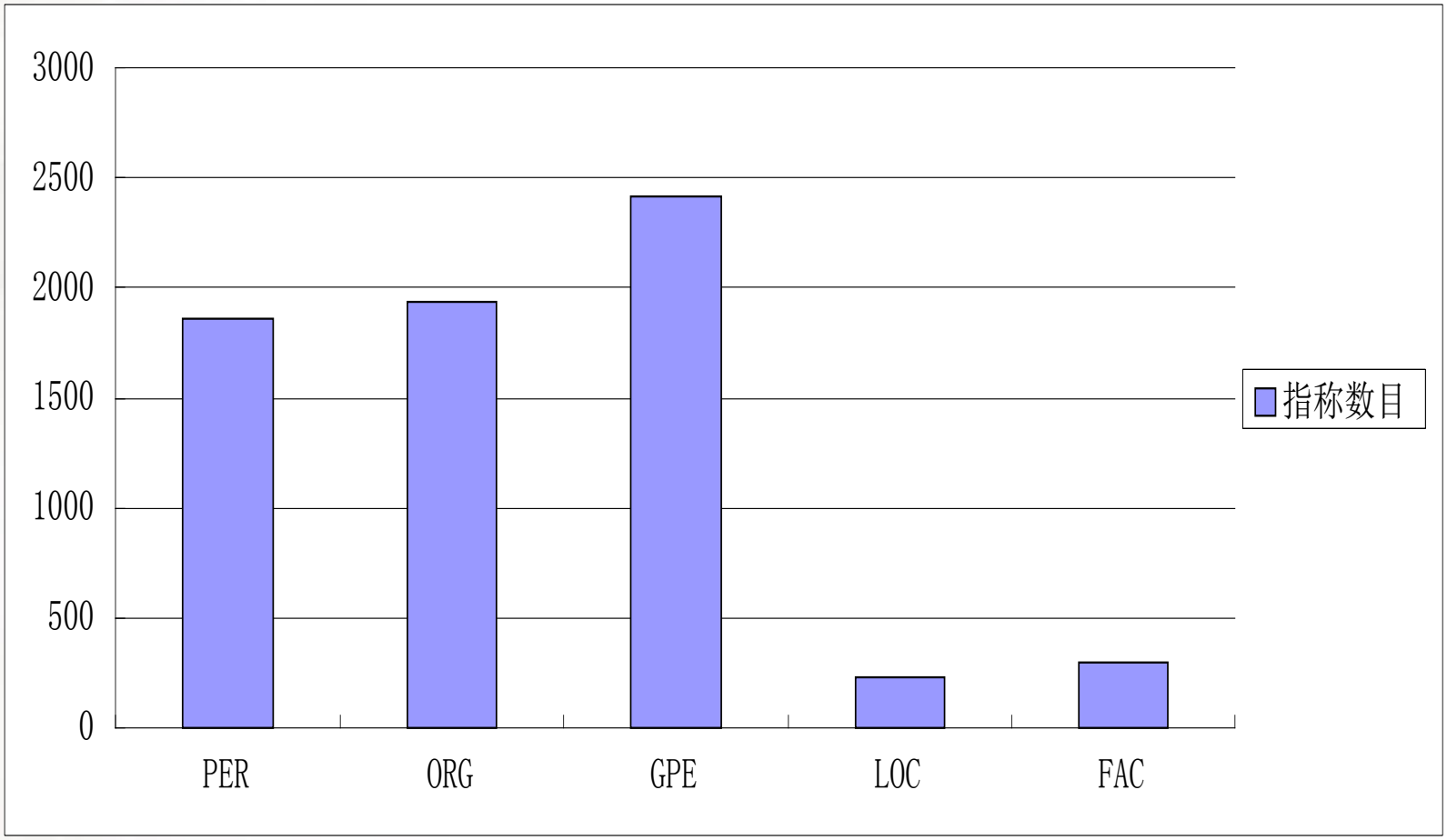
2).Human annotation

Because of the name variation and name duplication, the automatic annotation can not solve the all annotation, we need to modify the annotation by human to ensure the reliability.

# Annotation of the corpus

## 4.Statistics of our corpus

1.The number of different entity type in ACE2005c

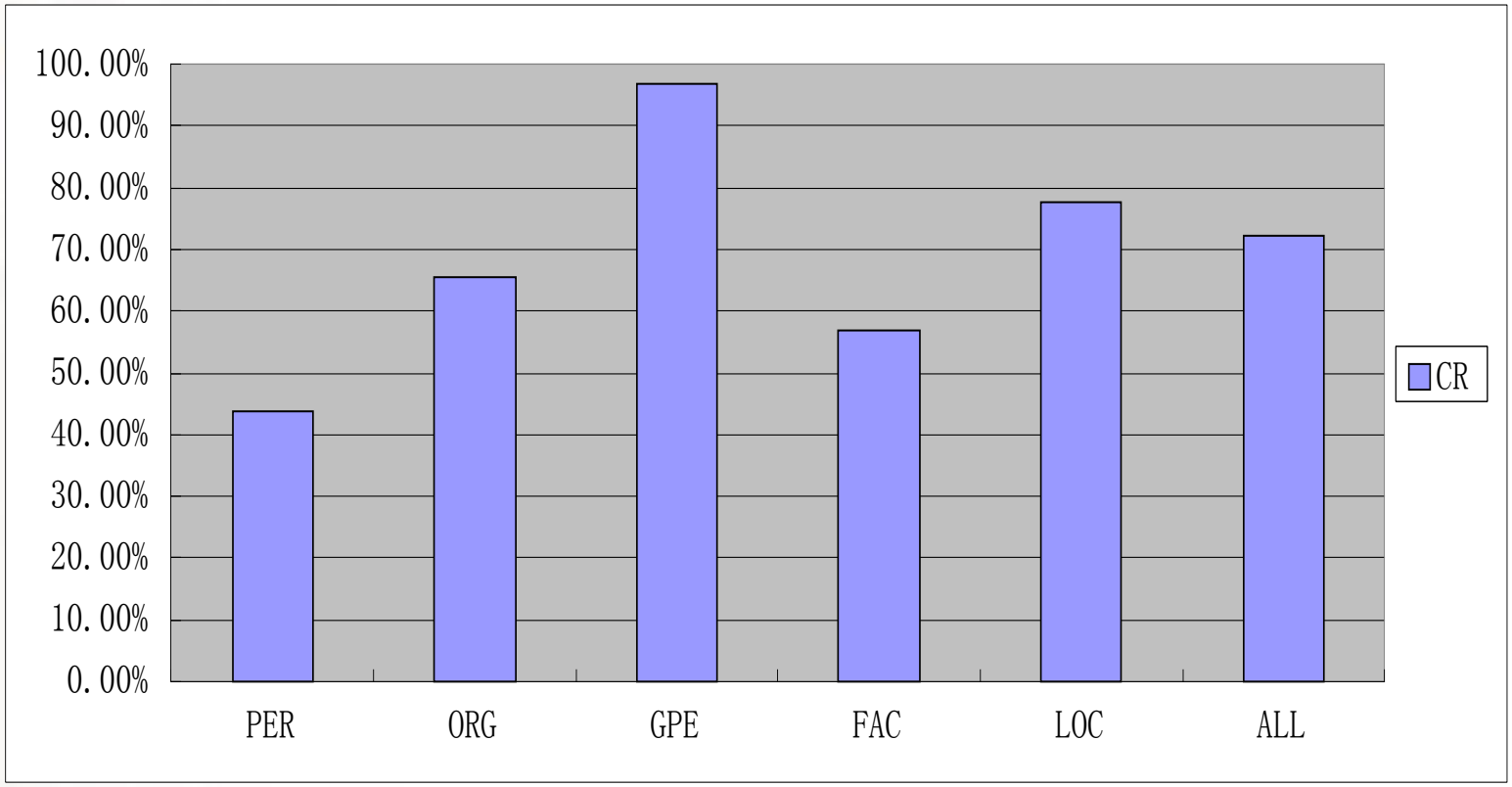


# Annotation of the corpus

If one entity mention can find corresponding entity in knowledge base, we think it is covered.

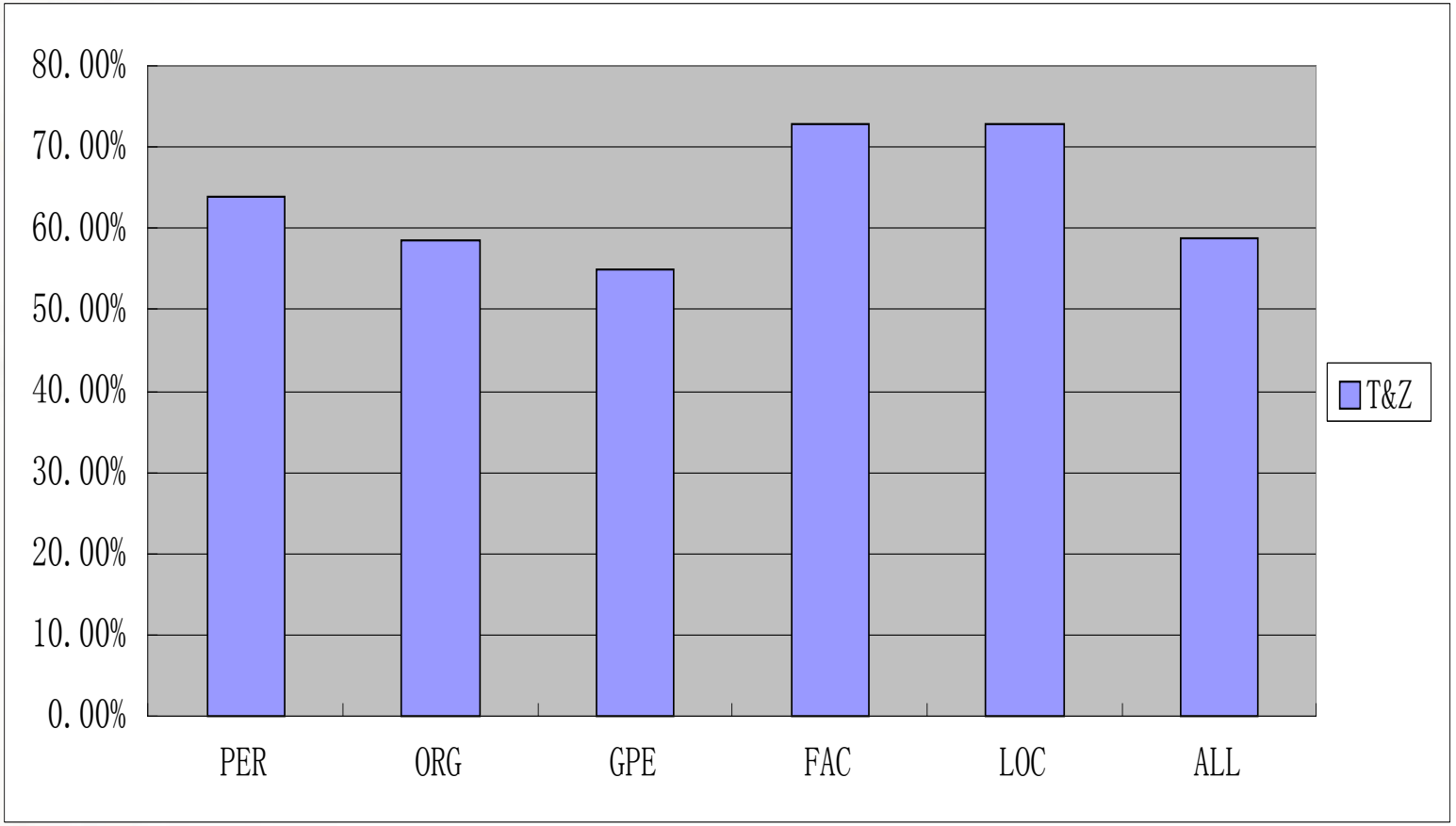
CR(Covered Ratio)=Covered/Total

2.The coverage of Wikipedia for the entity mentions in ACE2005c



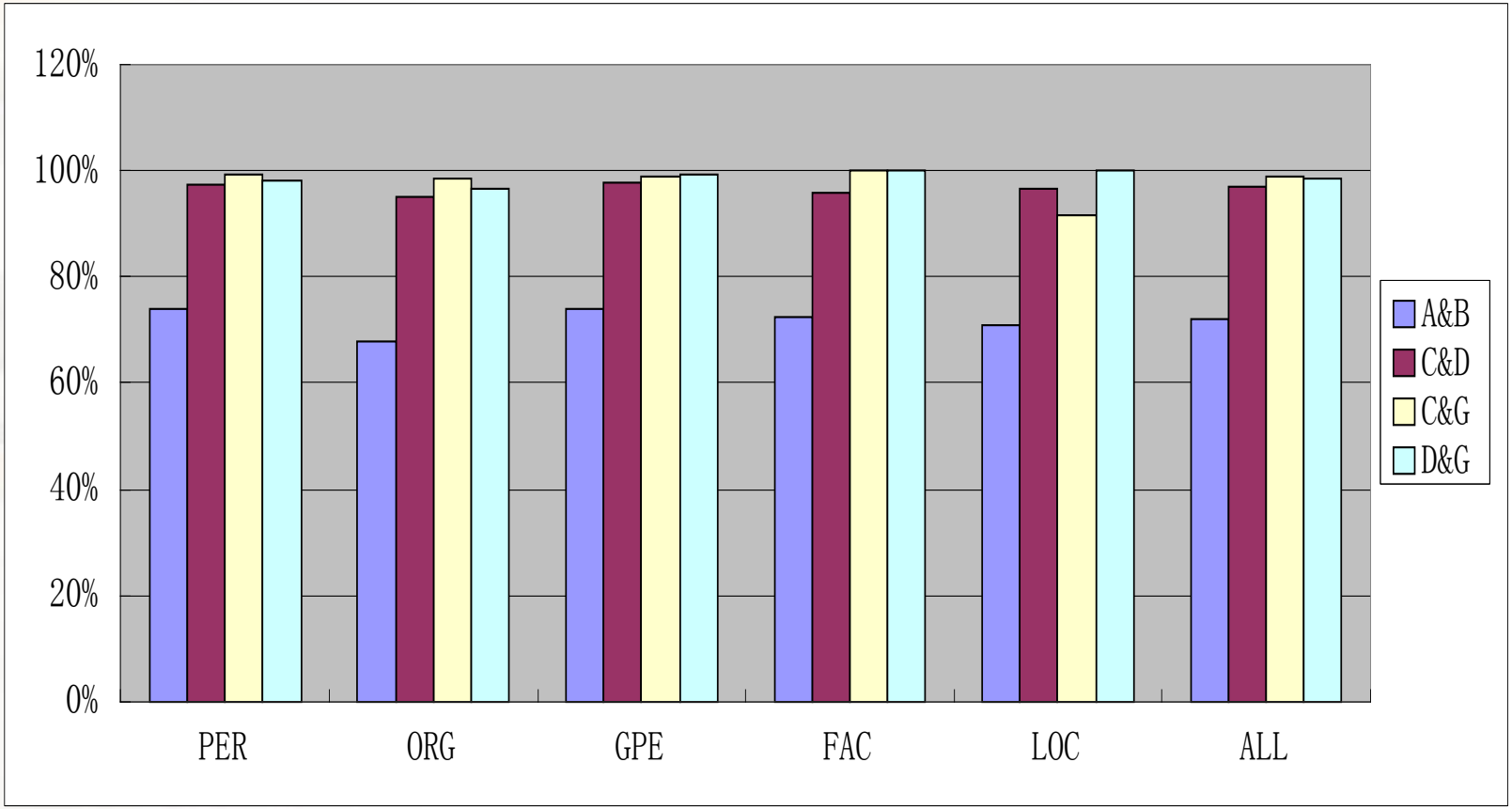
# Annotation of the corpus

3.The consistent rate of automatic annotation and final annotation.



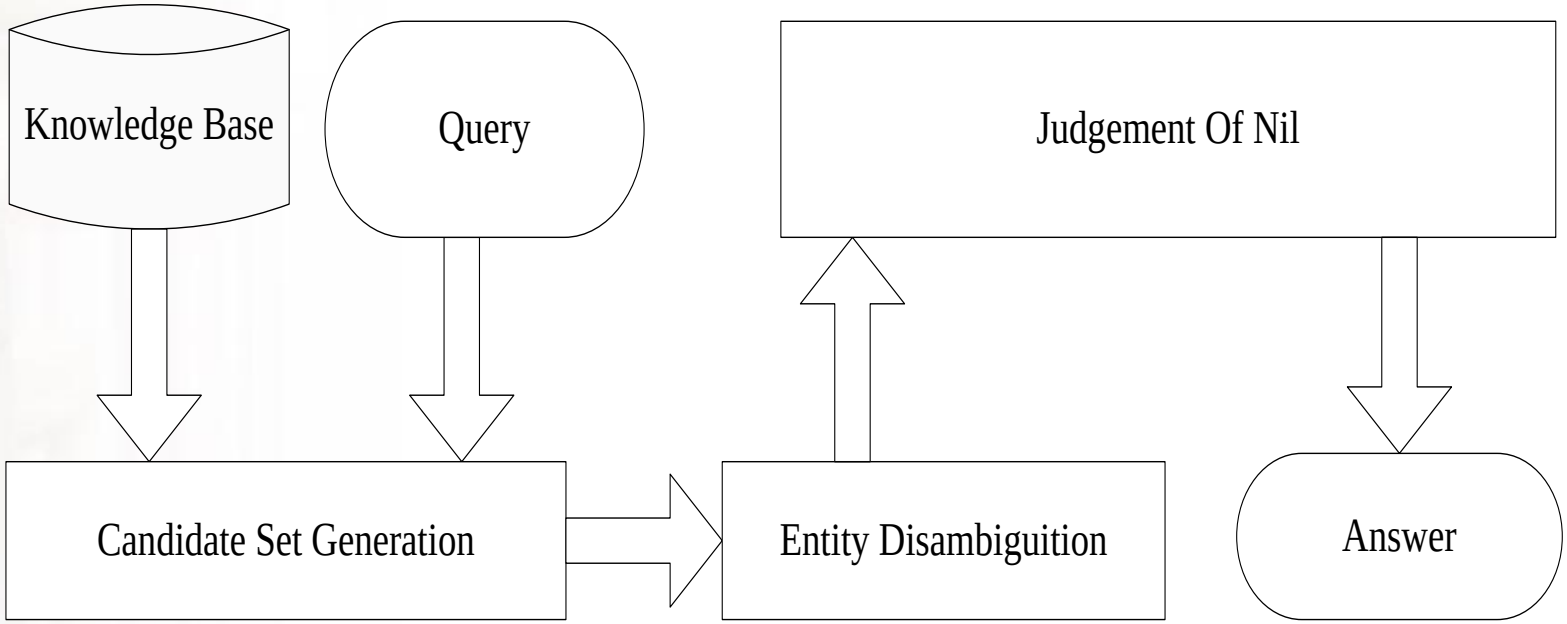
# Annotation of the corpus

4.Coincidence scores before and after adjustment



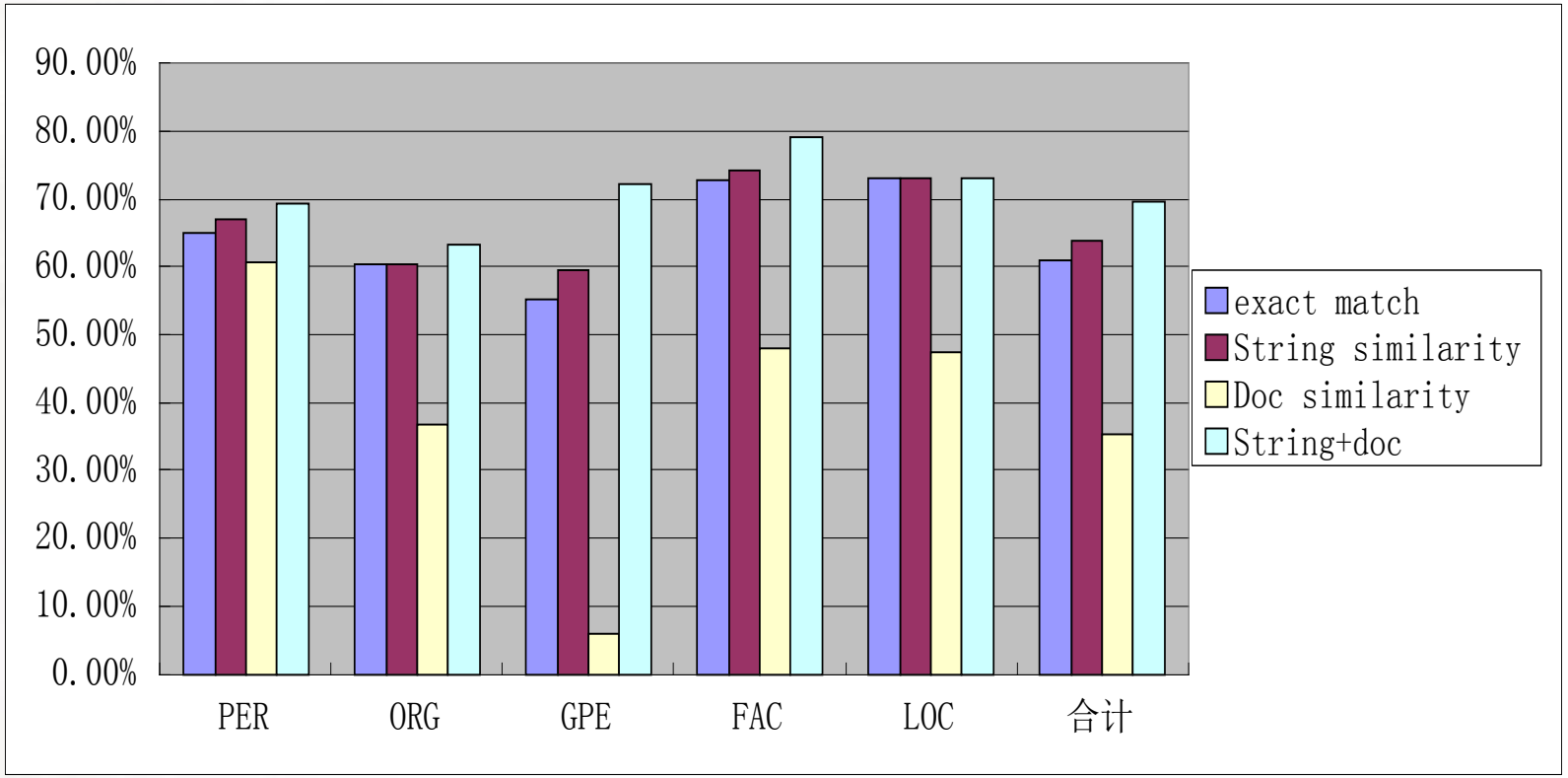
# Baseline

flowchart of our baseline



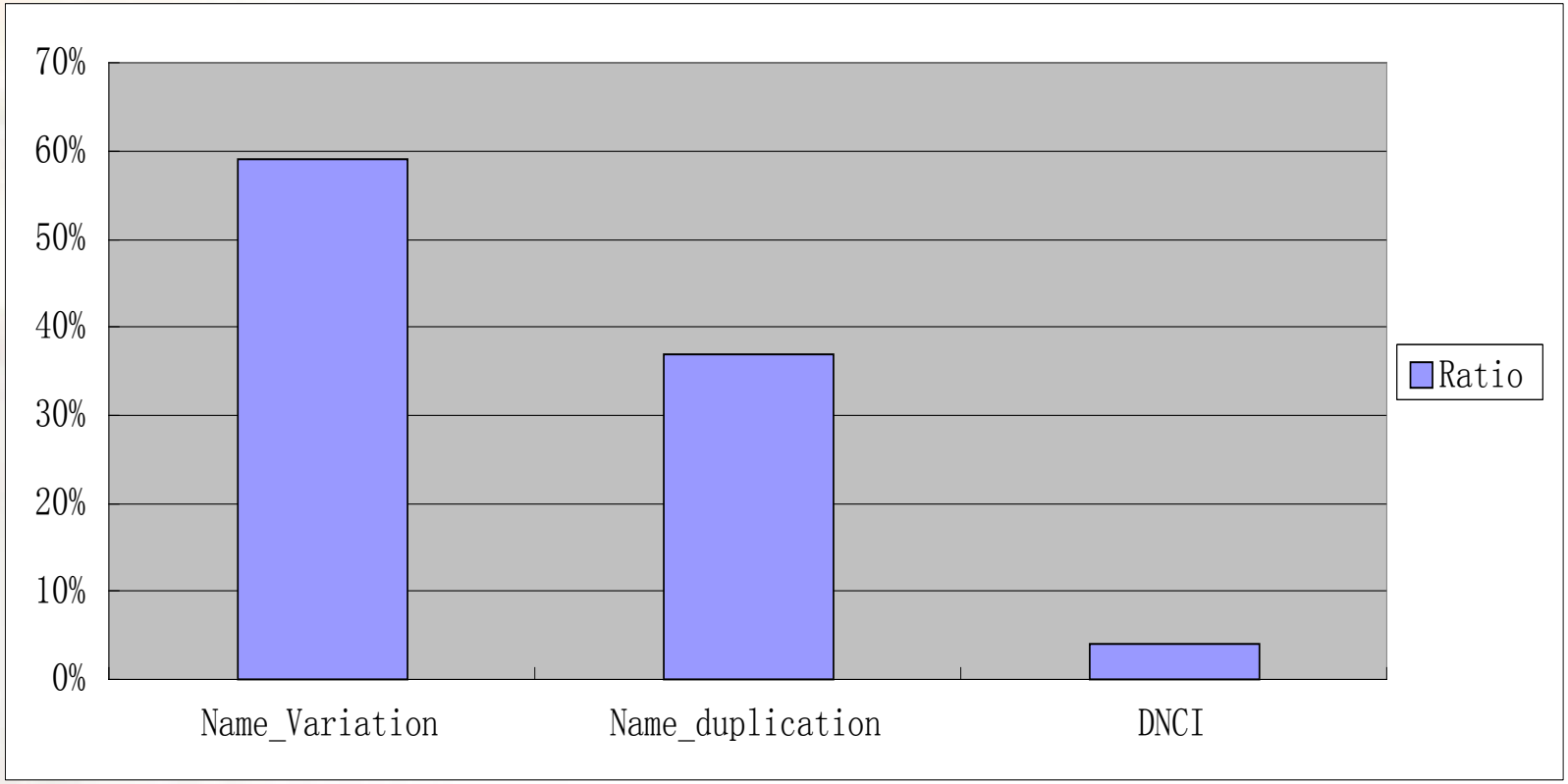
# Baseline

The accuracy of our baseline



# Baseline

## Analysis of errors




DNCI:Doc Not Contains Information

# Future and Prospect

The accuracy of our baseline and the analysis of errors of baseline inspire us that we should modify the method of candidate generation, mine more language features to help entity disambiguation, and improve the performance judgement of Nil.

# Comments and Question?

Thank you!