Research and Exploration of Entity Naming Standards

Authors: Liu Jianhua, Guo Hongmei, Liu Jianhua

Date: 2016-06-12T00:00:00+00:00

Abstract

This paper takes entity name normalization—one of the fundamental tasks in text processing—as its theme, and clarifies the twoAbstractThis paper takes entity name normalization—one of the fundamental tasks in text processing—as its theme, clarifies two types of tasks in entity name normalization: the entity coreference resolution problem where one entity has multiple names, and the entity disambiguation problem where one name refers to different entities. Combining these two types of tasks, it comprehensively analyzes current related research achievements, focuses on introducing typical ideas and methods for solving entity name normalization currently, important projects and significant evaluation conferences that promote entity name normalization research, and combined with existing problems in current research, analyzes and discusses the research trends of entity name normalization.

Full Text

Preamble

Study on Named Entity Normalization

Liu Jianhua[1],[2], Guo Hongmei[1],[2]

[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100190, China)

Abstract This paper focuses on Named Entity Normalization (NEN), a fundamental task in text processing. It delineates two core sub-tasks within NEN: entity co-reference resolution (where one entity has multiple names) and entity disambiguation (where one name refers to different entities). Synthesizing these two tasks, we comprehensively analyze current research findings, emphasizing

typical approaches and methodologies for solving NEN, key projects and evaluation conferences that drive NEN research forward, and discuss future research trends by examining persistent challenges in the field.

[**Keywords**] Named Entity Normalization; Entity Disambiguation; Large-scale Knowledge Base; Social Network

In the real world, different people often assign different names or descriptions to the same entity. With the continuous development of information technology and the proliferation of online resources, the diversity of entity names has grown substantially, posing significant challenges for automated computer understanding and computation. To support downstream text processing tasks such as machine translation, information retrieval, and data mining, it becomes necessary to map these varied names and descriptions to their corresponding entities and select a canonical representation as the core linkage among different expressions. This need has given rise to the concept of Named Entity Normalization.

From a thematic perspective, research closely related to Named Entity Normalization includes entity co-reference resolution, abbreviation recognition, and entity disambiguation, with corresponding English terms such as "Named Entity Disambiguation," "Abbreviation Reorganization," "Co-reference Resolution," and "Named Entity Normalization." From a task-oriented viewpoint, NEN encompasses two types of problems: (1) the entity co-reference problem where one entity has multiple names, which includes both pronominal anaphora resolution (e.g., identifying the referents of "he," "she") and nominal reference resolution (e.g., determining whether "44th President of US," "Barack Obama," and "President Obama" refer to the same person); and (2) the entity ambiguity problem where one name may refer to different entities [1]. Due to factors such as the enumerative nature of sense representation (from a limited set of meanings to rule-based generation of new senses), fine-grained sense distinctions (from subtle differences to antonyms), domain-specific versus loosely-defined natural text, a single entity name often corresponds to multiple named entity concepts. For instance, "Washington" may refer to either the state of Washington or the first U.S. president. Resolving such ambiguity requires identifying the specific concept intended.

This paper examines Named Entity Normalization as its central theme, synthesizing and analyzing current domestic and international research findings. We focus on typical approaches and methodologies for NEN, important projects and evaluation conferences that advance the field, and explore future research trends by analyzing existing challenges.

2 Main Approaches and Methods for Named Entity Normalization

Named Entity Normalization is fundamentally a computational process for automatically discerning the true meaning of terms in context [2]. While it shares

similarities with Word Sense Disambiguation, NEN is more complex due to the lack of comprehensive named entity concept lists and the greater diversity of entity mention forms (full names, abbreviations, aliases, pronouns, short forms, and cross-linguistic spelling variations such as British vs. American English). Accomplishing this task requires extensive knowledge, including not only common linguistic knowledge such as shallow lexical, syntactic, and grammatical analysis, but also semantic and background information. This paper surveys major current research and distills three mainstream methodological approaches.

2.1 Web Object Attribute-Based Approaches

Web pages often embed various objects such as person, product, and organization names. Extracting and integrating these objects from web pages enables powerful object-level content discovery. The advantage of this approach lies in the unique nature of its data source—web resources offer great convenience for obtaining object attributes, thereby facilitating attribute template-based coreference resolution.

Zaiqing Nie et al. [3] define Web objects as data units describing certain Web information, typically viewed as concepts related to application domains. A Web object can be represented through a series of attributes A = {a , a , ···, a }, where the attribute set can be predefined according to domain requirements. In practice, Nie et al. [4] refer to structured lists of similar items on the Web (such as product lists or service directories) as data records. Their approach first extracts domain-relevant data records from sources to create object record-level identifiers, then performs attribute-level extraction by analyzing these records to identify different parts as distinct attributes. By obtaining different attribute values for the same object from multiple sources, object fusion is ultimately achieved based on these attribute values.

While this method offers high convenience and accuracy, it has significant limitations, imposing strict requirements on data source formats and being applicable only to a small number of structured or semi-structured web pages that describe entities.

2.2 Large-Scale Knowledge Base-Based Approaches

The key challenge in entity disambiguation is measuring the similarity of entity mentions. Traditional methods employ Bag-of-Words (BOW) models, but these ignore semantic relationships. With the emergence of structured and semi-structured knowledge bases online, many researchers have proposed leveraging resources like Wikipedia [5] and YAGO [6] to construct large-scale knowledge bases, using the background knowledge they provide to improve NEN effectiveness. This has become a central focus in current NEN research.

Wikipedia is often the first choice for researchers due to its broad concept coverage, with each article containing information about an entity or concept, rich semantic information, and continuously updated content. Anthony Fader et

al. [7] introduced the GROUNDER system, which effectively utilizes prior information from Wikipedia' s user-contributed content and novel disambiguation models, combining prior and contextual information to improve disambiguation accuracy. Hien T. Nguyen et al. [8] map entities mentioned in text to correct Wikipedia entries, demonstrating that combining Wikipedia and textual features yields optimal disambiguation results based on a candidate entity statistical ranking model. Danuta Ploch [9] frames NEN as the task of linking entity mentions in text to predefined referents in a knowledge base, mining co-occurrence relationships between entities in Wikipedia to derive classification features for candidate entities and combining disambiguation functions using SVM classifiers to achieve effective results.

However, since Wikipedia itself has limitations in data accuracy and concept structure representation, many researchers have turned to Linked Open Data (LOD), which offers stronger advantages in accuracy and relationship expression through manual curation and organization. Danica Damljanovic et al. [10] argue that Linked Data is an effective resource for expanding available context, combining advanced named entity tools with LOD-based similarity measures to demonstrate improved Wikipedia disambiguation accuracy. Kamel Nebhi [10] employs FreeBase combined with syntactic parsing for word sense disambiguation, with experiments showing improved performance.

Beyond LOD, ontologies with richer semantic-level associations have also become important knowledge bases for NEN exploration. Horacio Saggion et al. [11] conducted research on cross-data source knowledge unit acquisition and integration based on the EU MUSING platform, dividing the process into ontology-based information extraction and ontology-based cross-data source object integration. A key feature of their system is a business ontology constructed by domain experts, containing class hierarchies, relationships, and attributes in the business domain, with defined objects including company names, employee counts, addresses, websites, phone numbers, fax numbers, and profit status. After annotating each document to obtain annotated objects and their descriptive content, similarity is computed to cluster identical objects from multiple data sources, thereby achieving named entity normalization. Farhad Abedini [12] utilizes the vast number of factual descriptions between entities provided by YAGO to identify semantic entities in text. Xianpei Han et al. [13] comprehensively leverage multiple knowledge sources including WordNet, Wikipedia, and web information to mine contextual semantic information of entity mentions, proposing a graph-based knowledge representation model that integrates heterogeneous semantic information within a unified framework to mine potential semantic associations between concepts, thereby effectively integrating semantic knowledge from different sources and improving NEN efficiency.

2.3 Social Network-Based Approaches

With the continuous development of search engines and social network mining technologies, leveraging social relationships to construct social networks for en-

tity resolution has emerged as a key approach. These methods are primarily applied to person name disambiguation, typically using spectral clustering to cluster names in social networks, then introducing modularity thresholds as stopping conditions for network partitioning based on the influence of edge weights and graph partition criteria on disambiguation effectiveness [14].

[Figure 1: see original paper] illustrates a typical social network-based NEN framework. In this area, Ron Bekkerman et al. [15] proposed an unsupervised framework to address the problem of retrieving numerous irrelevant person pages when searching for a specific individual. Two key components are web page link relationships and Agglomerative duplicate clustering, where link relationships are used to construct a person' s social network. Lang Jun et al. [16] exploit the principle that different individuals sharing the same name have different social networks, using co-occurring person names in search results to discover and expand potential social networks related to the query person. Combining graph spectral partitioning algorithms with modularity metrics for automatic social network clustering, they achieve disambiguation of search results for ambiguous names. Experiments on manually annotated Chinese person name corpora demonstrate good overall performance, with graph clustering algorithms helping to further partition connected social networks and improve disambiguation effectiveness.

Chen Chen et al. [14] first use spectral clustering to cluster person names in social networks, then introduce modularity thresholds as stopping conditions for network partitioning based on the impact of edge weights and graph partition criteria on name disambiguation effectiveness, achieving good results in co-reference resolution. Jaderick P. Pabico [17] addresses entity name ambiguity in social networks by proposing a graph-subgraph approach to determine similarity between different entities. Mohammad et al. also propose constructing co-authorship networks using heuristic clustering methods to resolve author name ambiguity caused by integrating multi-source data in digital libraries.

3 Key Projects and Evaluation Conferences for Named Entity Normalization

Research on Named Entity Normalization has been driven forward by major projects and international evaluation conferences. This section surveys these important initiatives to provide reference for future researchers.

3.1 Major NEN Projects

(1) UK National Archives TNA-Search Project [18]

The UK National Archives (TNA) represents a landmark project in large-scale entity name normalization. As part of the Government Web Archive Project, TNA-Search aims to develop simple, intuitive mechanisms to improve access to government website records dating back to 1997, comprising approximately 700

million web pages. To address NEN challenges, the project primarily utilizes GATE integrated with FactForge and SKB (Semantic Knowledge Base) Ontology to construct a large-scale semantic repository (Large Knowledge Base, LKB). This repository provides detailed object descriptions and background information to compute and achieve entity name normalization.

Specifically, the project directly associates entities in documents with various ontologies through the LKB, either via instances or concepts. The LKB uses a series of SPARQL query configuration files to retrieve information from SKB. Entity annotation and SKB instance association are accomplished through two complementary approaches: when a match is found via the LKB dictionary, SKB class and instance information is added to relevant entities in the text; when no direct association exists between text entities and SKB classes or entities, association is achieved through co-reference. That is, if a mention in the text has already been associated with SKB, all co-referential mentions can automatically obtain the same class and instance information through the TNA Instance Generator. During canonical annotation, the project associates different expressions of the same entity within a document while adding feature relationships between annotations discovered by the semantic tagger. Through this approach, TNA-Search achieves automatic annotation and normalization of 11 types of named entities including persons, geographical names, organizations, and temporal expressions.

(2) OKKAM [19]

OKKAM is a large-scale integration project funded under the EU' s Seventh Framework Programme (FP7). Based on the 14th-century "Occam' s razor" principle, it advocates not increasing entity identifiers unless necessary. OKKAM provides a global infrastructure for content creators, editors, and developers called the Entity Name System (ENS), which includes a feature-based instance matching method (FBEM). FBEM identifies potential object co-references by integrating similarity measures across multiple different feature attributes and their values from two instance identifiers. For example, FBEM uses a Levenshtein edit distance-based method to compare local names of instance identifiers.

(3) Domestic Projects

Co-reference resolution and entity disambiguation are crucial tasks in text processing that significantly improve information retrieval efficiency and enable deep text mining. China has several relevant research projects in this area, notably Tsinghua University' s RiMoM [20] and Nanjing University' s ObjectCoref [21].

RiMoM is a multi-strategy ontology matching system developed by Tsinghua University that integrates various ontology matching methods, including multiple instance matching approaches. For instance matching, RiMoM categorizes instance information into six types: URL, metadata, name, string-type information, non-string-type information, and neighbor information. It computes similarity between instances using edit distance-based methods and vector space

models, filters results using metadata and non-string-type information, and finally integrates various similarity measures through multiple strategies to identify object co-references.

Unlike RiMoM, Nanjing University' s ObjectCoref is built on datasets provided by the semantic web search system Falcons, currently containing over 73 million instance identifiers. ObjectCoref first constructs an initial training set through semantic equivalence reasoning, including owl:sameAs, functional or inverse-functional properties, and cardinality restrictions. It then continuously learns from this training set in a bootstrapping manner to identify object co-references. A key technique is learning the most suitable attributes and attribute values for identifying co-references from the training set. The system also considers frequent attribute combinations, using two attributes simultaneously to identify co-references (e.g., longitude and latitude, first name and last name) to further improve accuracy. Additionally, it ranks instance identifiers that co-refer to the same object based on whether semantic equivalence relations are dereferenceable and the frequency of instance identifier occurrences across different RDF documents. ObjectCoref proposes a novel architecture integrating semantic equivalence reasoning with similarity computation that can comprehensively identify object co-references, though erroneous co-reference relationships in the training set may lead to error accumulation during learning and reduced accuracy.

3.2 NEN Evaluation Conferences

Numerous international evaluation conferences have promoted the continuous development of NEN research by refining evaluation tasks, providing corpora, and offering communication platforms. This section introduces several typical evaluation conferences for reference.

(1) Automatic Content Extraction (ACE) and Text Analysis Conference (TAC)

The ACE initiative began in July 1999 and officially launched in December 2000, co-sponsored by the NSA, NIST, and CIA, with eight sessions held to date [22]. ACE evaluation tasks include Entity Detection and Recognition (EDR), Value Detection and Recognition (VAL), Time Expression Recognition and Normalization (TERN), Relation Detection and Recognition (RDR), and Event Detection and Recognition (VDR). Co-reference resolution evaluation is primarily embedded within EDR, which maps various mentions in documents to corresponding entities to provide comprehensive entity descriptions. This task requires first identifying all mentions, then merging those describing the same entity—a process that constitutes co-reference resolution. Notably, ACE has included Chinese language evaluations since 2003, with five sessions conducted to date, representing the only international evaluation for Chinese co-reference resolution.

After 2008, ACE was succeeded by the Text Analysis Conference (TAC) [23].

TAC-KBP has been held six times since 2009, with the Entity Linking task being directly related to NEN. The current TAC Entity Linking task uses a target entity knowledge base built from the October 2008 version of Wikipedia, containing nearly 820,000 entities: 110,000 person entities, 55,000 organization entities, 110,000 geographical entities, and 530,000 entities of other categories, with a total knowledge base size of approximately 2.6GB [24].

(2) Web People Search Evaluation (WePS)

WePS is a specialized evaluation conference focusing on person name disambiguation in English web pages, organized primarily by Julio Gonzalo and Satoshi Sekine, with three sessions held to date [25]. The task concentrates on person name disambiguation in web search scenarios. Participating systems receive a person name as a search query and must determine how many distinct individuals appear in the search results and assign specific mentions to the correct individual. Overall, this is a clustering problem: given a set of documents, cluster them according to the real-world person referred to by a specified name, ensuring all mentions in each cluster refer to the same person. WePS3' s evaluation tasks emphasize person attributes including birth date, birthplace, aliases, occupation, affiliation, awards, education, degree, major, ethnicity, phone number, and temporal information. Inspired by this project, Wenjie Li et al. organized a Chinese person name disambiguation evaluation task in 2010 [26], with two sessions held to date.

(3) Anaphora Resolution Exercise (ARE) [27]

From November 2006 to March 2007, the University of Wolverhampton initiated the Anaphora Resolution Exercise (ARE), the most comprehensive co-reference resolution evaluation conducted in English to date, comprising four tasks: - Pronoun resolution on pre-annotated documents: All noun phrases are identified and pronouns requiring resolution are marked; systems must find correct antecedents for each pronoun from a list of noun phrases not containing pronouns. - Co-reference resolution on pre-annotated documents: All noun phrases are identified; systems must recognize all co-reference chains within documents. - Pronoun resolution on raw text: Unlike the first task, documents contain no annotations; systems must identify relevant information themselves. - Co-reference resolution on raw text: Unlike the second task, documents contain no annotations; systems must identify relevant information themselves.

In addition to these domain-independent evaluations, there are domain-specific co-reference resolution tasks such as JNLPBA (Joint Workshop on Natural Language Processing in Biomedicine and Its Applications) and BioCreAtIvE (Critical Assessment of Information Extraction Systems in Biology). These conferences continuously advance NEN research.

4 Research Trends in Named Entity Normalization

Despite mature development of NEN research, evaluation results (e.g., average system efficiency of 72.1% in TAC 2012 entity linking [28]) indicate that current recognition performance remains insufficient for large-scale practical applications, with many challenges yet to be addressed, including nil entity problems, knowledge base coverage issues, knowledge base inaccuracy, and knowledge base utilization problems [29]. Consequently, research in this field is evolving along several key trends.

(1) Trend Toward Multi-Model Fusion

In past research, statistical methods based on linguistic features and machine learning approaches were considered separately, with many studies incorporating linguistic features only as selected features in classification or clustering. This limited fusion approach yielded modest improvements in recognition efficiency. Current research increasingly considers using linguistic insights to construct richer machine learning models. Elango [30] proposed an initialization approach combining centering theory with Conditional Random Fields (CRF) for pronoun resolution. The flexibility of CRF models effectively incorporates context-dependent preference transmission. Poesio et al. [31] treat clauses as discourse units, representing documents as sequences of clauses and thus as feature spaces composed of lists of forward-looking centers. This feature space can integrate relevant features such as grammatical roles, gender, and number. Similar sequential CRF model inference and estimation can employ techniques discussed by Sutton and McCallum [32].

(2) Increasingly Diverse Feature Selection for Disambiguation

Current research publications show growing emphasis on introducing increasingly diverse features into NEN, with "knowledge-poor" approaches that rely solely on algorithmic improvements falling out of mainstream favor. Commonly used entity disambiguation features are summarized in .

** Summary of Entity Disambiguation Features**

Feature Category	Specific Features
Morphological/Syntactic	Number, distance, person, string matching, part-of-speech
Syntactic	Dependency relations, grammatical roles
Semantic	Entity type, entity attributes (varying by entity type), apposition, aliases, Wikipedia category overlap
Contextual	Intra-sentence entity co-occurrence, contextual similarity, Wikipedia article in-link and out-link text

The growing number of applied features is supported by the emergence of various corpus resources that provide access to deep linguistic knowledge through three primary channels: (1) Conventional knowledge bases such as WordNet, HowNet, Wikipedia, DBPedia, and YAGO; (2) Pattern mining from large-scale corpora, such as Hearst' s [33] "is-a" templates for discovering synonyms from text, Bergsma' s [34] extraction of English noun phrase gender and number information from Minipar-parsed corpora, and Yang and Su' s [35] use of template information discovered in corpora to enhance co-reference resolution; (3) Utilizing the entire Internet as a corpus, using search engine result counts to compute various relevance measures, such as Poesio et al.' s use of mutual information to examine associations between phrases.

(3) Automatic Construction of Large-Scale Knowledge Bases as a Critical Component

Experiments consistently demonstrate that high-quality large-scale knowledge bases strongly support improved NEN efficiency. Faced with exponentially growing web data, manual expert construction of knowledge bases is clearly time-consuming and leads to information lag. Therefore, automatic construction of large-scale knowledge bases rich in semantic associations is particularly crucial. Research on open information extraction and the emergence of large-scale semi-structured web knowledge bases like Wikipedia and Freebase provide a solid foundation for this endeavor. Representative work includes YAGO, which stores knowledge as instance-relationship triples derived from Wikipedia category pages and linked to WordNet, with confidence annotations for each factual entity achieving 95% accuracy. YAGO2 contains 10 million entities and 120 million factual records describing entity relationships [6]. Additionally, Zhao Jun et al. at the Institute of Automation, Chinese Academy of Sciences, leveraged information extraction techniques to expand an encyclopedia knowledge base from 80,000 to millions of entries by using the "Encyclopedia of China" knowledge system as the target structure, extracting concept instances from web knowledge bases, and integrating semantic tags, semi-structured information, and unstructured information from encyclopedia pages. This provides knowledge resource support for developing open-domain question answering systems.

This paper has conducted an extensive analysis of theories and methods related to Named Entity Normalization, examining mainstream approaches, typical domestic and international projects, and evaluation conferences to provide deep understanding of NEN' s core content. By analyzing practical challenges facing NEN research, we have explored future trends in this evolving field.

[1] Hien T. Nguyen, Tru H. Cao. A Knowledge-Based Approach to Named Entity Disambiguation in News Articles[J]. AI 2007, LNAI 4830, pp. 619–624, 2007

[2] ROBERTO NAVIGLI. Word Sense Disambiguation: A Survey[J]. ACM Computing Surveys, Vol. 41, No. 2:

[3] Nie, Z., et al., Web object retrieval[C]. In: Proceedings of the 16th international conference on World Wide Web, 2007: 81-90

[4] Nie, Z., et al., Object-level ranking: bringing order to Web objects[C]. In: Proceedings of the 14th international conference on World Wide Web, 2005: 567-574
[5] Wikipedia. http://www.wikipedia.org/[EB/OL] (Accepted: 2014-11-26)
[6] YAGO2s: A High-Quality Knowledge Base. http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/[EB/OL] (Accepted: 2014-11-26)
[7] Fader, A., Soderland, S., Etzioni, O.: Scaling Wikipedia-based named entity disambiguation to arbitrary web text[C]. In: Proceedings of the IJCAI Workshop on User-contributed Knowledge and Artificial Intelligence: An Evolving Synergy, Pasadena, CA, USA, pp. 21-26 (2009)
[8] Hien T. Nguyen, Tru H. Cao Exploring Wikipedia and Text Features for Named Entity Disambiguation[J]. Intelligent Information and Database Systems Lecture Notes in Computer Science, 2010, Volume 5991/2010: 11-20
[9] Danuta Ploch. Exploring Entity Relations for Named Entity Disambiguation[C]. In: Proceedings of the ACL 2011 Student Session, Portland, OR, USA, 2011
[10] Danica Damljanovic, Kalina Bontcheva. Named Entity Disambiguation using Linked Data[EB/OL]. http://2012.eswc-conferences.org/sites/default/files/eswc2012_submission_334.pdf (Accepted: 2014-11-26)
[11] Horacio Saggion, Adam Funk, Diana Maynard, Kalina Bontcheva. Ontology-based Information Extraction for Business Intelligence[EB/OL]. https://gate.ac.uk/sale/iswc07/musing/musing-iswc07.pdf (Accepted: 2014-11-26)
[12] Kamel Nebhi. 2013. Named entity disambiguation using freebase and syntactic parsing[C]. In: Proceedings of the First International Workshop on Linked Data for Information Extraction (LD4IE 2013) co-located with the 12th International Semantic Web Conference (ISWC 2013)
[13] Xianpei Han. Named Entity Disambiguation by Leveraging Wikipedia Semantic Knowledge[C]. In: Proceedings of the 18th ACM conference on Information and knowledge management, 2009: 215-224
[14] 陈晨，王厚峰. 基于社会网络的跨文本同名消歧 [J]. 中文信息学报, 2011(5): 75-82
[15] Ron Bekkerman, McCallum Andrew. Disambiguating web appearance of people in a social network[C]. In WWW ' 05 Proceedings of the 14th international conference on World Wide Web, 2005: 463-470
[16] 郎君，秦兵等. 基于社会网络的人名检索结果重名消解 [J]. 计算机学报, 2009(7): 1-10
[17] JADERICK P. PABICO. An Analysis of Named Entity Disambiguation in Social Networks[J]. Asia Pacific Journal of Multidisciplinary Research, 2014(2): 31-38
[18] Diana Maynard, Mark A. Greenwood. Large Scale Semantic Annotation, Indexing and Search at the National Archives[EB/OL]. https://gate.ac.uk/sale/lrec2012/tna/tna.pdf (Accepted: 2014-11-26)
[19] Paolo Bouquet, Themis Palpanas, Heiko Stoermer, Massimiliano Vignolo. A Conceptual Model for a Web-scale Entity Name System[EB/OL]. http://www.inf.unibz.it/krdb/events/swap2010/paper-19.pdf (Accepted: 2014-11-26)

[**20**] Li JZ, Tang J, Li Y, Luo Q. RiMOM: A dynamic multistrategy ontology alignment framework. IEEE Trans. on Knowledge and Data Engineering, 2009, 21(8): 1218−1232
[**21**] ObjectCoref. http://ws.nju.edu.cn/objectcoref/[EB/OL]. (Accepted: 2014-11-26)
[**22**] Automatic Content Extraction (ACE) Evaluation. http://www.itl.nist.gov/iad/mig/tests/ace/[EB/OL]. (Accepted: 2014-11-26)
[**23**] Text Analysis Conference. http://www.nist.gov/tac/[EB/OL]. (Accepted: 2014-11-26)
[**24**] KBP 2013 Entity Linking Task Description V1.0. http://www.nist.gov/tac/2013/KBP/EntityLinking/guid (Accepted: 2014-11-26)
[**25**] Javier Artiles, Andrew Borthwick, Julio Gonzalo, Satoshi Sekine, Enrique Amigo. WePS-3 Evaluation Campaign: Overview of the Web People Search Clustering and Attribute Extraction Tasks[EB/OL]. (Accepted: 2014-11-26)
[**26**] CLP2012-Chinese Language Processing. http://www.cipsc.org.cn/clp2012/bakeoff-cn.html[EB/OL]. (Accepted: 2014-11-26)
[**27**] Constantin Orasan, Dan Cristea, Ruslan Mitkov, Antonio Branco, Anaphora Resolution Exercise: An overview. http://www.lrec-conf.org/proceedings/lrec2008/pdf/713_paper.p (Accepted: 2014-11-26)
[**28**] Jeffrey Dalton, Laura Dietz. A Neighborhood Relevance Model for Entity Linking. http://ciir.cs.umass.edu/~dietz/entitylinking/oair2013.pdf[EB/OL]. (Accepted: 2014-11-26)
[**29**] 赵军，刘康，周光有，蔡黎. 开放式文本信息抽取 [J]. 中文信息学报, 2011(6): 98-110
[**30**] P. Elango. Coreference resolution: A survey. Project report of the course "Advanced natural language processing" [D], In computer science departments, university of Wisconsin Madison, 2006
[**31**] M. Poesio, M. Kabadjov. A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation[C]. In: Proc. of the 4th International Conference on Language Resources and Evaluation. Lisbon, Portugal
[**32**] C. Sutton, A. McCallum. 2006. An introduction to conditional random fields for relational learning[C], In: L. Getoor and B. Taskar, eds. Introduction to statistical relational learning: MIT Press
[**33**] M.A. Hearst. Automatic acquisition of hyponyms from large text corpora.[C]. In: Proceedings of the 14th International Conference on Computational Linguistics, 1992
[**34**] S. Bergsma. Automatic acquisition of gender information for anaphora resolution[C]. In: B. Kégl and G. Lapalme eds. Canadian Conference on AI, 2005, Victoria, Canada: Springer-Verlag, 342-353
[**35**] X. Yang, J. Su. Coreference resolution using semantic relatedness information from automatically discovered patterns[C]. In: J. Carroll, A. Bosch, and A. Zaenen eds. Proc. of the 45th Annual Meeting of the Association of Computational Linguistics. Prague, Czech Republic: Association for Computational Linguistics, 528-535

Author Biographies:

Liu Jianhua (1984-), female, from Nantong, Jiangsu, intermediate professional, Ph.D. candidate, primarily engaged in text mining and information extraction research.

Guo Hongmei (1985-), female, from Zhoukou, Henan, Ph.D. candidate, primarily engaged in text mining and scientometrics research.

Note: Figure translations are in progress. See original paper for figures.