An Integrated Approach to Heterogeneous Data for Information Extraction

Ying Chen, Sophia Y. M. Lee, and Chu-Ren Huang

Department of Chinese & Bilingual Studies,
The Hong Kong Polytechnic University
{chenying3176, sophiaym, churenhuang}@gmail.com

Abstract. The paper proposes an integrated framework for web personal information extraction, such as biographical information and occupation, and those kinds of information are necessary to further construct a social network (a kind of semantic web) for a person. As web data is heterogeneous in nature, most of IE systems, regardless of named entity recognition (NER) or relation detection and recognition (RDR) systems, fail to get reliably robust results. We propose a flexible framework, which can effectively complement state-of-the-art statistical IE systems with rule-based IE systems for web data, and achieves substantial improvement over other existing systems. In particular, in our current experiment, both the rule-based IE system, which is designed according to some web specific expression patterns, and the statistical IE systems, which are developed for some homogeneous corpora, are sensitive only to specific information types. Hence we argue that our system performance can be incrementally improved when new and effective IE systems are added into our framework.

Keywords: relation extraction, information extraction

1 Introduction

Semantic web, which collects and formats different kinds of web knowledge, plays an important role in the development of a new generation of web. One important component of semantic web is to automatically extract different relations existing in web data. Information extraction (IE) can provide such a technology to solve this problem, particularly for a specific named entity. For example, Web People Search[1] (WePS) 2009 evaluation (Sekine & Artiles, 2009) tries to extract some personal information, and TREC Entity Track[2] plans to find some related information for a product.

Web IE is particularly challenging because web data is heterogeneous in nature. Complete or comprehensive IE information necessarily come from many different sources with different formats. For example, “affiliation” and “email” are so different in their own expressions so that they need different extraction approaches. Hence, a homogeneous IE model, regardless of statistical model or rule-based model, often cannot perform effectively for web IE. To overcome this problem, some previous systems (Culotta et al., 2004; Lan et al., 2009; Watanabe et al., 2009) have tried to combine different IE approaches, which are often homogeneous IE, to extract different types of information in web data. Nevertheless, few of them have explored how to effectively utilize or integrate different IE tools for web data.

[1] http://nlp.uned.es/weps/

[2] http://ilps.science.uva.nl/trec-entity/

In this paper, we propose a framework to integrate heterogeneous IE approaches for web IE. In this framework, we first segment web data according to the expression format. Similar to the genre categories – "formal text" and "informal text" – which were defined in Minkov et al. (2005), a text is either in "formal style" or "informal style." A "formal style" text obeys prescribed writing standards, i.e. a complete sentence usually with a subject and an object. On the contrary, "informal style" has few limitations on writing format and can mix various representation levels. In order to do so, we develop a novel algorithm to segment a webpage into fragments according to their expression format: formal style and informal style. This segmentation allows an existing IE system, which often was developed for a specific type of text, to be applied to its similar text fragments. For example, most statistical IE systems are developed for a news corpus, therefore it is better to apply them to formal-style fragments.

In addition, web data also have their ways of conveying certain information. For example, it is common that occupation and affiliation information is expressed in a homepage in the format of "Name, Position, Affiliation," such as "Anita Coleman, Assistant Professor, University of Arizona." As this kind of web-specific expression is often multi-lines, some existing IE patterns (Mann & Yarowsky, 2003; Mann, 2006; Rosenfeld & Feldman, 2006), which were limited to one sentence or were designed for formal style text, cannot be directly applied. To identify this web expression property, we develop web-specific patterns, which take into account of different kinds of information, such as webpage type information (i.e. homepage and biographical webpage), and text expression style (formal style and informal style). The experiment shows that those patterns could achieve high precision, which is very important for real applications.

Instead of presenting a totally new IE solution for web data, the goal of this paper aims at providing a flexible framework, which is able to effectively reuse and integrate different existing well-developed IE technologies for web data, and tries to collect some web-specific information to further help IE. We test our IE framework on a small scale data provided by the WePS 2009 evaluation[3], one of whose tasks is to extract some personal information for a given person, and the experiment shows a promising result. Moreover, the comparatively-high precision of our system indicates the strong capability of our framework to integrate IE technologies. Finally, according to the personal information distribution in web data, we discuss the practical problems of IE systems for further improvement. In this paper, we concentrate only on IE for a focus person. The terms "attribute" and "relation" are interchangeable here.

2 Related work

Although IE is an old topic, it still poses a big challenge, especially for web data. In general, IE contains two key components: named-entity recognition (NER) and relation detection and recognition (RDR). In the personal IE case, NER, which extracts possible attribute value candidates, is a basic component, and RDR, which detects relations involving the focus person and given attribute value candidates and further selects valid attribute values for that person. To effectively tackle IE, both NER and RDR are required to have a good performance.

For NER, the naïve approach is rule-based, but its big disadvantage is the difficulty of rule-design (Feldman, 2002) or rule-learning (Etzioni et al., 2005), which needs to handle various named-entity expressions. In recent years, statistical NER technology (Bikel et al., 1999; McCallum & Li, 2003) has been significantly improved through a series of evaluations, such as Automatic Content Extraction (ACE), Message Understanding Conference (MUC), and so on. However, because those NER technologies mainly focused on news documents, it was so dependent on text information, such as capitalization information and corpus type, that their performance dropped much when working on web data because those cues are rather noisy.

[3] http://nlp.uned.es/weps/

Besides, the adaptation of these NER systems to other kinds of corpus is not easy (Vilain et al., 2007). Although some NER systems, e.g. Minkov et al. (2005), attempted to do the adaptation work from news corpus to a non-news corpus, they still focused on a homogeneous corpus. Nevertheless, web data is heterogeneous in nature and it is impossible to know the source information of the documents. Therefore, it is very difficult to do NER adaptation for web data. In this paper, we explore a problem: how to effectively re-use the existing well-developed NER system for web data.

Compared to NER, RDR is still a comparatively hot and difficult topic. Although some statistical RDR systems have been developed, such as the systems participating ACE, they were usually designed only for homogeneous data as most of the existing statistical NER systems. Therefore, most of the previous web text mining systems adopted rule-based approach to extract information (Rosenfeld et al., 2004; Soderland, 1999). Similar to rule-based NER, the main problem of rule-based RDR is the difficulty in designing rules for all kinds of text. In recent years, some studies have been done to learn rules by a semi-supervised or totally unsupervised approach (Mann & Yarowsky, 2003; Mann, 2006; Rosenfeld & Feldman, 2006). These approaches only detect relations existing in a sentence, which is not enough for web data as some relations occur across sentences. Therefore, some patterns specific to web data needs to be learned. Overall, RDR is still on the stage of exploration now.

Most of the previous work has put much effort to develop a statistical IE system mainly focusing on a homogeneous corpus, in particular on news corpus, and their adaptation to web data is not an easy task. In this paper, we adopt another approach to solve web IE: how to effectively integrate those existing IE systems for web data. Meanwhile, we also explore some web-specific expression patterns in web personal information expressions.

3 Methodology

Our IE framework consists of two main components: preprocessing (webpage type detection and fragment segmentation) and personal information extraction. Preprocessing is very important in our framework as it allows our system to integrate different IE technologies for personal information extraction.

3.1 Preprocessing

Given a webpage, it is first categorized into three webpage types according to its relationship to the focus person, namely homepage, related webpage (a webpage mainly describes the focus person, such as biographical webpage), and others. It is then segmented into several fragments based on its text expression styles, which could be formal style fragment or informal style fragment. Figure 1 gives an example of the two fragments expressing the similar information. Formal style fragment gives information in a complete sentence in a conservative manner. For informal style fragment, it gives only keywords, usually with each piece of information in a separate line. Keywords are often capitalized.

Formal style fragment
Anita Sundaram Coleman is an Assistant Professor in the School of Information Resources & Library Science at the University of Arizona, Tucson, which she joined in 2001.
Informal style fragment
Anita Coleman
Assistant Professor
School of Information Resources & Library Science
1515 E. First St.
University of Arizona
Tucson, AZ 85719

Figure 1: Examples of two kinds of fragments

```
if   title of a webpage contains a keyword for "homepage"
        web type =  "homepage"
elif   title is the person name:
        web type =  "homepage"
elif   title contains the person name
        web type =  "related page"
elif   the URL of a webpage contains the last name or the first name of the personal name:
        web type = "homepage"
else
        web type = "others"
```

Figure 2: The algorithm for webpage type

3.1.1. Webpage type detection

The type of a webpage can sometimes provide important document-level information for IE. For example, all occurrences of “I” in a homepage refer to the focus person, and therefore all information in those sentences is about the focus person. It is not easy to completely catch the type information of a webpage because a webpage creator may put this information in various places. In this study, we apply some naïve rules only to a webpage title and its URL to detect its webpage type. The details of the rule are presented in Figure 2.

3.1.2. Text fragment segmentation

It is common that a webpage is often written in a mixture of different representations: formal style and informal style. For example, in a resume, the description of “objective” is often in formal style, while the “education experience” section is more likely to be in informal style. This noisy structure of webpage brings a lot of trouble to IE processing, no matter using rule-based or machine-learning (ML)-based approaches.

As mentioned, most of the current ML-based IE systems were trained from a corpus, whose expression format is similar to formal style, and cannot be effectively applied to informal style text. To reuse these well-developed ML-based IE systems, we first segment a webpage into fragments, and then apply a ML-based IE system only to those formal style fragments. Another advantage of our fragment segmentation is that a fragment is often a comparatively small unit, so it becomes much easier to design rules just focusing on one fragment.

There are two steps in our fragment segmentation. First, each line in a webpage is classified as one of the two classes – formal style or informal style – according to the percentage of tokens that begin with capitalization, as it is assumed that informal style text mainly consists of capitalized words. Second, continuous lines that share the same expression type are considered as a single fragment. For instance, consider a 10-line webpage as below:

Line 1: *********** formal
Line 2: *********** formal
Line 3: *********** formal
Line 4: *********** informal
Line 5: *********** formal
Line 6: *********** informal
Line 7: *********** informal
Line 8: *********** formal
Line 9: *********** informal
Line 10: ********** informal

Each line is classified as either formal or informal style. Lines 1, 2 and 3 are linked as one fragment, which is followed by the other five fragments, i.e. Line 4, Line 5, Line 6-7, Line 8, and Line 9-10. There are six fragments in total.

3.2 Personal information extraction

As explained, the final IE result is decided by the total performances of both NER and RDR in this IE system. In this paper, we mainly focus on how to effectively combine and reuse different NER and RDR technologies for web data. Currently, we explore the two main categories of IE technologies: rule-based and ML-based. According to the combination ways of NER and RDR systems, we have three types of IE systems: a pure rule-based IE, a pure ML-based IE, and a hybrid IE (consisting of a ML-based NER and a rule-based RDR.)

For rule-based IE, we develop rule-based NER and RDR systems especially for web data, and for ML-based IE, we adopt some existing ML-based NER and RDR systems, which were developed for news corpora, to handle web data. Our rule-based NER and RDR systems differ from previous rule-based IE with the limited scope of rules and the consideration of some web-specific information. All rules are designed limited to a fragment so that it can save a lot of effort to find a rule that can effectively work in a whole document. Meanwhile, we also take some web-specific information into account when designing rules. In the following section, we first briefly describe each NER and RDR system involving in our IE systems, and then give the three types of personal IE systems.

Table 1: List of attributes in the WePS 2009 AE task

Attribute names					
date of birth	birth place	other name	occupation	affiliation	relatives
phone	fax	email	website	nationality	
degree	major	school	mentor	award	

3.2.1. IE components.

Rule-based NER: Since informal style text is often noisy, almost no existing ML-based NER system is suitable for this kind of text. Here, we develop a rule-based NER system to extract the 16 kinds of attributes (listed in Table 1) used in WePS 2009 task. The rules are all tailor-made for formal and informal style text, and each attribute has different rules.

First, a keyword set is collected for each attribute in question. For example, the "occupation" keyword set is taken from "Dictionary of Occupational Titles" (DOT), and the "organization" keyword set from General Architecture for Text Engineering (GATE)[4]. Then, some patterns, which use the keyword sets, are used to extract named entity expressions in question.

ML-based NER - BBN IdentiFinder: Compared to informal-style fragment, formal-style fragment is well-written, and thus many well-developed ML-based NER systems can directly be applied to formal fragments and a fairly good performance can be achieved. In our experiment, we choose *BBN IdentiFinder*, which has a nice user interface and has a reasonable performance.

[4] http://gate.ac.uk/

Rule-based RDR: Given the named-entity expressions detected by a NER system, no matter it is rule-based or ML-based, their relationships with the focus person need to be detected by a RDR system. There are two kinds of patterns in our rule-based RDR system: keyword-based pattern, and web-specific pattern.

A keyword pattern is similar to previous patterns for text mining which identify relations within a sentence. First, a keyword set is collected for each target attribute. For example, the keyword "born" is chosen for the attributes of "date of birth" and "birth place." Then, a relation between a named entity expression and a focus person exists only when the following two requirements are satisfied.

1) The named entity expression belongs to the required types, which is defined by the target attribute. For example, for the attribute "occupation," the named entity must be an organization.

2) A keyword for the target relation must appear in that sentence.

Besides the keyword-based patterns, we also design some patterns, which can search for personal information in question in the whole fragment so as to identify some web-specific expressions, i.e., the pattern in the informal style example in Figure 1. The "occupation" and "affiliation" attributes in that example are listed one by one in a fragment. Currently, we tried to catch some common attribute-listed patterns in webpages, especially in homepages, for personal information expression.

ML-based IE - EXERT: *EXERT* (Hacioglu & Chen, 2005) is a complete IE system, which was developed for the ACE 2005 project, whose corpus is mainly a news collection. It includes three components: NER, co-reference and RDR. The *EXERT* system achieved comparative performances for all of the three components, especially in RDR, in the ACE 2005 evaluation.

3.2.2. Personal IE systems.

According to different combination ways of the IE components given above, we develop three personal IE systems, and each one represents a type of integrated methods.

Rule-based IE: The IE system consists of the rule-based NER and the rule-based RDR.

Hybrid IE: The IE system includes *BBN IdentiFinder* and the rule-based RDR. To achieve a high precision, this hybrid IE system is applied only to five attributes: date of birth, birth place, occupation, affiliation and school.

ML-based IE: it is a direct application of the *EXERT* system to the WePS 2009 task. Because of the different relations, on which the WePS and ACE projects focus, we make a mapping to convert the ACE output format to the WePS personal attribute output format. For example, if the type of the mention of the focus person is "nominal" in the ACE output, the mention string is assigned as "occupation" in the WePS output. However, the EXERT system provides only six WePS attributes (relations): nationality, affiliation, school, relatives, occupation and other name.

Table 2: Focus fragments for each IE system

	Homepage and related page	Others
Rule-based	Any fragment	Any fragments containing a mention of the focus personal name.
Hybrid-based	Any formal style fragment	The sentences in formal style fragments containing a mention of the focus personal name.
ML-based	The sentences in formal style fragments containing a mention of the focus personal name.	The sentences in formal style fragments containing a mention of the focus personal name.

As mentioned, different IE components are often designed or developed for a specific type of text. To save time and achieve high accuracy, for each IE system and each webpage, considering the information gotten from preprocessing, we can choose only some specific fragments or sentences (as listed in Table 2) to run it. In our framework, first, all sentences that contains the various expressions of the focus personal names, i.e., "Anita S. Coleman," "Coleman, Anita" for "Anita Coleman," are detected using the rules in our rule-based NER. Then, each IE system runs separately, but all of them are limited to the text according to Table 2.

4 Experiment

In this paper, we make use of the WePS 2009 corpus to conduct the experiment. The Web People Search (WePS) evaluation provides a forum for a standard evaluation, which focuses on IE for personal named-entities in web data. There are two tasks in the WePS 2009 evaluation: clustering and attribute extraction (AE). The clustering task, which can also be called personal name disambiguation, groups those webpages according to whether the given personal name occurring in that webpage refers to the same person in reality. Attribute extraction, which can be considered as a special case of IE, extracts certain personal information for a focus person with the given personal name. In this paper, we focus only on the AE task. Although the WePS 2009 corpus is a small scale corpus comparing to the huge web data, but it is still able to show the web personal information distribution, and to prove how our IE framework works for web data.

4.1 Data Analysis

The WePS 2009 AE corpus includes 18 personal names in the training data and 30 personal names in the test data, and there are totally 3,468 documents (Sekine & Artiles, 2009). The data set for each personal name consists of about 100 webpages that contain the focus personal name. For each webpage, 16 kinds of attributes (referring to Table 1) could be extracted if existing.

First, we want to get some ideas about personal information distribution in web data, which can reflect the reasonability of our text-expression-format division (formal and information fragments). For each webpage in the WePS 2009 AE test data, for each golden-standard attribute value in that webpage, we search the webpage to catch the occurring frequency of this attribute value in a formal fragment and in an informal fragment, and show in Table 3. Notice, some attribute values may occur in both formal and information fragments.

In Table 3, we can see that the distribution varies depending on attribute types. In general, some simple attributes, whose values often can be caught by some fixed patterns, such as"email," "phone," and "fax," are more likely to be expressed in an informal style, whereas

Table 3: The personal information distribution in the WePS 2009 test data

	affiliation	occupation	birth place	date of birth
Formal-style	5,010	7,595	347	251
Informal-style	3,672	4,344	190	265
	email	fax	phone	website
Formal-style	67	3	50	84
Informal-style	163	68	217	79
	degree	major	school	mentor
Formal-style	340	165	478	551
Informal-style	519	135	397	164
	other name	award	nationality	relatives
Formal-style	1,182	211	665	1,591
Informal-style	739	116	224	492

some complicated attributes, whose values often are extracted by a ML-based NER, such as “relatives,” “affiliation,” and “occupation,” are more likely to occur in formal-style fragment. Nevertheless, some attributes, such as “date of birth,” “school”, do not show any preference in their expression ways in web data. Table 3 can also give some idea to improve personal IE system, as it indicates which attributes tend to be expressed in formal or informal text styles. In such a way, we will know where more effort is needed. For example, for “email,” it is better to do more work on web-specific patterns, whereas for “occupation,” a traditional ML-based NER and RDR may be a good choice.

4.2 Performances

All of our experiments run on the WePS 2009 AE test data, and the results are evaluated by the scoring provided by WePS 2009. We first run the pure rule-based IE system, and then add the hybrid IE system and the pure ML-based IE system one by one. The performances are presented in Table 4. In addition, we also give the performance of purely ML-based IE system.

In Table 4, we notice that performance is consistently increasing when incorporating more IE technologies, and either rule-based IE or ML-based IE cannot work well for web data. The final combination system (Rule-based + Hybrid IE + ML-based IE), which can complement both rule-based and ML-based technologies, has achieved the best F score (18.89). It beats the top system in the WePS 2009 AE evaluation (Sekine & Artiles, 2009): 12.22. However, the low F score also indicates that personal information extraction from web data is still a big challenge. Therefore, more effort is needed in this respect.

In addition, comparing to most AE systems participating in the WePS 2009 AE evaluation, our system has achieved a higher precision. One possible cause of the phenomenon of high recall and low precision in these systems is the noisy NER systems they used to detect possible attribute value candidates. However, in our system, first we notice that the rule-based IE system has very high precision (36.31), which is much higher than the highest precision (30.4) reported in the WePS 2009 AE evaluation (Sekine & Artiles, 2009). This indicates our web-specific patterns are effective for web IE. Meanwhile, high precision is very important for real applications. Moreover, we also find that incorporating ML-based IE can improve recall while not hurting precision too much. This indicates that our integrated approach, which is based on the web type information and the web text expression style, can effectively combine the two different NER technologies (rule-based and ML-based) into one system for web data.

We also show the detailed performances (F scores) of the 16 kinds of WePS attributes in Table 5. As mentioned, both the hybrid IE system and the ML-based IE system affect the extraction of several attributes, so the performances of other attributes do not differ when incorporating those two IE systems. Therefore, we use “same” in Table 5 to indicate no change.

It is not surprising to notice, in Table 5, that the attributes of “email,” “fax,” and “phone” has achieved very good performances even with the rule-based IE, because they are almost fixed expressions. On the other hand, the attributes of “affiliation,” “occupation,” and “award” do not perform well even with the final combination system, because they can be expressed in various ways, which cannot be easily caught either by the rule-based system or by the ML-based IE system. Nevertheless, from Table 5, we can notice that integrating of different IE can complement the performances for some specific attributes, and further improve the overall performance. When we look at the performance variation closely, we found that the increase of the overall performance is largely due to the improvement of the precision, which indicates that our framework can compatibly integrate different IE technologies for web data, and therefore is flexible to add more IE components.

As shown in Table 5, we find that the performances of “birth place” and “date of birth” improve significantly after incorporating hybrid NER. However, in Table 3, we know that “birth” information appears almost evenly in both kinds of fragment, especially for “date of birth.” This indicates the birth information in formal style is easier to be detected if the birth value candidates can be detected by a NER, whereas this kind of information in informal style is somewhat noisy. When adding the ML-based IE system, we find the performances of “affiliation” and “occupation” has improved, and this phenomenon is consistent with the information distribution (the information of “affiliation” and “occupation” is more likely in formal style text.) The performances of “nationality” and “relatives” do not change much although the ML-based IE should be able to extract this kind of information. This indicates this kind of information extraction needs more effort.

Table 4: Performances of our IE systems on the WePS 2009 test data

	precision	recall	F score
Rule-based IE	36.31	9.15	14.62
ML-based IE	28.95	5.19	8.80
Rule-based + Hybrid IE	37.06	10.91	16.86
Rule-based + Hybrid IE + ML-based IE	31.90	13.42	18.89

Table 5: Performances for each attribute of our IE systems on the WePS 2009 test data (“same” means this extraction approach is the same as the previous one)

	affiliation	occupation	birth place	date of birth
Rule IE	5.9	12.6	15.2	11.6
Rule + Hybrid IE	6.2	14.9	38.6	31.7
Rule + Hybrid + ML IE	9.1	20.5	same	same
	email	fax	phone	website
Rule IE	43.0	51.5	35.0	17.0
Rule + Hybrid IE	same	same	same	same
Rule + Hybrid + ML IE	same	same	same	same
	degree	major	school	mentor
Rule IE	31.0	4.3	12.1	0.6
Rule + Hybrid IE	same	same	12.4	same
Rule + Hybrid + ML IE	same	same	12.3	same
	other name	award	nationality	relatives
Rule IE	37.4	13.2	18.0	2.0
Rule + Hybrid IE	same	same	same	same
Rule + Hybrid + ML IE	32.3	same	18.0	2.4

5 Conclusion

In this paper, we present a framework, which integrates heterogeneous IE approaches for web personal information extraction. The small-scale experiments presented in this paper shows our framework is able to flexibly combine the rule-based and ML-based NER for web personal IE, and the results are promising. In addition, heuristic patterns are developed to effectively catch the web information from heterogeneous sources. These patterns can also be added incrementally to improve the performance. Hence, we believe that the present framework is a very robust approach to heterogeneous information extraction.

It is important to note that, compared with uniform statistical systems, our integrated system has very high precision yet lower recall. This is because each integrated information approach covers only a percentage of the heterogeneous texts, yet extracted exactly the right target information. This is one reason why we believe our integrated system can be incrementally improved when new patterns or technologies are incorporated. While improvements with a uniform model will come at a much higher cost.

Nonetheless, the problem of web personal information extraction is far from being solved and more work is needed. We find that ML-based NER can improve the recall, so using a high-quality NER system for formal style text is one of our further work. However, it is still a big challenge to develop a high-quality NER just for informal style text in web data because of noisy surface cues. Moreover, the personal information distribution suggests that many complicated relations are expressed in formal style, and our fragment segmentation allows the use of existing RDR systems developed for formal style text. Therefore, we need to incorporate ML-based RDR, which can detect more kinds of attributes, into our system in the future. Finally, the question as to how to effectively extract and use web-specific information in personal information extraction, such as webpage type, web-specific patterns and so on, also needs more exploration.

References

Bikel, D. M., R. L. Schwartz, and R. M. Weischedel. 1999. An algorithm that learns what's in a name. *Machine Learning*.

Culotta, A., R. Bekkerman, and A. McCallum. 2004. Extracting social networks and contact information from email and the web. In *Proceedings of CEAS-04*.

Etzioni, O., M. Cafarella, D. Downey, A-M Shaked, S. Soderland, D. S.Weld, and A. Yates. 2005. Unsupervised named-entity extraction from the Web: An experimental study. *Artificial Intelligence, 165*, 91-134.

Feldman, R. 2002. Text mining. In: Kloesgen W, Zytkow J (eds.) *Handbook of Data Mining and Knowledge Discovery*. MIT Press, Cambridge, MA.

Hacioglu, K. and Y. Chen. 2005. University of Colorado (CU) ACE 2005 System, *ACE-05 Evaluation Workshop*, NIST, Gaithersburg, MD.

Lan, M., Y. Z. Zhang, Y. Lu, J. Su, and C. L. Tan. 2009. Which Who are They? People Attribute Extraction and Disambiguation in Web Search Results. In *2nd Web People Search Evaluation Workshop (WePS 2009)*, 18th WWW Conference.

Mann, G. and D. Yarowsky. 2003. Unsupervised Personal Name Disambiguation. In *Proceedings of CoNLL-2003*.

Mann, G. 2006. *Multi-Document Statistical Fact Extraction and Fusion*, Ph.D. Thesis.

McCallum, A. and W. Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of CoNLL-2003*.

Minkov, E., R. Wang, and W. Cohen. 2005. Extracting Personal Names from Emails: Applying Named Entity Recognition to Informal Text. *HLT/EMNLP*.

Rosenfeld, B. and R. Feldman. 2006. URES : an Unsupervised Web Relation Extraction System. In *Proceedings of COLING/ACL*.

Rosenfeld, B., R. Feldman, and M. Fresko. 2004. TEG: a hybrid approach to information extraction. In *Proceedings of CIKM*.

Sekine, S. and J. Artiles. 2009. WePS 2 Evaluation Campaign: Overview of the Web People Search Attribute Extraction Task. In *2nd Web People Search Evaluation Workshop (WePS 2009)*, 18th WWW Conference, 2009.

Soderland, S. 1999. Learning Information Extraction Rules for Semi-Structured and Free Text. *Machine Learning, 34(1-3)*: 233-272.

Vilain, M., J. Su, and S. Lubar. 2007. Entity Extraction is a Boring Solved Problem - or is it?. *Proceedings of NAACL HLT*.

Watanabe, K., D. Bollegala, Y. Matsuo, and M. Ishizuka. 2009. A Two-Step Approach to Extracting Attributes for People on the Web in Web Search Results. In *2nd Web People Search Evaluation Workshop (WePS 2009)*, 18th WWW Conference.