Task Definition of Attribute Extraction Subtask for WePS-2

1. Overview

This subtask is to extract 18 kinds of "attribute values" of target individuals whose names appear on each of the provided Web pages. The organizer will distribute the target Web pages in their original format, (i.e., html), and the participants will be expected to extract attribute values from each page. The individual name associated with a particular page will be given, and the attribute values for that person should be extracted. Web pages containing multiple individuals sharing the same name will NOT be given.

All attributes to be extracted are listed in Table 1 below. Although there are 18 attributes listed, "Work" and "Location" will NOT be evaluated for WePS-2, since after annotating the 300 sample texts, those attributes were found to be very ambiguous due to the degree of variation for each of the target individuals.

	Attribute Class	**Examples of Attribute Value**
1	Date of birth	4 February 1888
2	Birth place	Brookline, Massachusetts
3	Other name	JFK
4	Occupation	Politician
5	Affiliation	University of California, Los Angeles
6	Work	The Secrets of Droon
7	Award	Pulitzer Prize
8	School	Stanford University
9	Major	Mathematics
10	Degree	Ph.D.
11	Mentor	Tony Visconti
12	Location	London
13	Nationality	American
14	Relatives	Jacqueline Bouvier
15	Phone	+1 (111) 111-1111
16	FAX	(111) 111-1111
17	Email	xxx@yyy.com
18	Web site	http://nlp.cs.nyu.edu

Table 1 Definition of 18 attributes of Person at WePS-2

In the following section, the general rules of the task will be explained. Section 3 includes a description of the ignored pages which will not be used in this evaluation. Section 4 provides participants with a detailed definition of each attribute as well as an explanation of potentially ambiguous cases. Section 5 provides an explanation of the evaluation metric.

2. General Rules

a) Attribute values should only be extracted from the pages provided. Those should be extracted AS IS. Attribute values which don't exist in the given pages should not be extracted. Do not extract a value from any pages that are linked from the pages provided.

b) If there are two or more attribute values for one attribute class, participants should extract all the values. For example, both "Japan" and "Tokyo" can be extracted as values of "Birthplace" from the expression, “He was born in Japan and the city of Tokyo.” However, if the two values are used in a single phrase, they must be extracted as one value. For example, the entire phrase “Tokyo, Japan” must be extracted from the expression, “He was born in Tokyo, Japan.”

c) An expression should be extracted even if it contains a factual error. For example, both “1782” and “June 25, 1841” should be extracted as values for “Date of Birth” from the following sentence: “Macomb, Alexander (1782-1841) General: Alexander Macomb was born on Detroit, Michigan, on June 25, 1841.”

d) If a single attribute value is expressed in more than one way on the same page, each expression should be extracted as a separate value.

Expression	**Attribute Value**
New York University (NYU)*****	New York University
	NYU
1949	1949
‘49	‘49
General	General
Gen.	Gen.
GENERAL	GENERAL
writer	writer
author	author
novelist	novelist
comic novelist	comic novelist
Stern School, NYU	Stern School, NYU
NYU, Stern School	NYU, Stern School
Stern School, New York University (NYU)	Stern School, New York University
	NYU
New York University (NYU), Stern School	New York University (NYU), Stern School

***** If an abbreviation appears in the middle of an extractable expression, do not extract the abbreviation as a separate value.

Expression	**Correct**	**Incorrect**
New York University (NYU), Stern School	New York University (NYU), Stern School	New York University, Stern School
		NYU

If an abbreviation is written at the end of the expression to be extracted, the abbreviation should be extracted separately, even if it appears as a part of the expression.

Expression	**Correct**	**Incorrect**
Stern School, New York University (NYU)	Stern School, New York University	New York University, Stern School (NYU)
	NYU	

The above rule does not apply to a postal address.

Expression:
3801 Bellemeade Avenue,
Evansville, Indiana (IN)

The above address should be extracted as a one expression. “IN” should not be extracted separately.

e) Do not extract a value written in a non-English language.

Expression	Correct	Incorrect
PhD (Doctorat)	PhD	PhD
		Doctorat

f) If a single attribute value appears differently because of spacing and/or punctuation (i.e., capitalization at the beginning of a sentence), it is not necessary to extract each expression. No penalty or advantage will be given to participants who produce the same value multiple times.

Expression	Attribute Value
New York University New York University	New York University
Tony Abbott is a writer… Writer Tony Abbott is…	Writer

g) If there is a line break in an attribute value, the break and spaces adjacent to the break can be replaced by a single space. No penalty or advantage will be given either way. The evaluation program will replace a sequence of spaces and breaks into a single space before the evaluation.

Expression:

715 Broadway, 7th floor
New York, NY 10003
USA

Can be left as it is (including line breaks) or extracted as:

715 Broadway, 7th floor New York, NY 10003 USA

h) The determiner ("the") at the beginning of a name is optional in the evaluation. No penalty or advantage will be given if the determiner is included or omitted.

Expression	Correct	Correct
The Beatles	The Beatles	Beatles
The University of Vermont	The University of Vermont	University of Vermont

i) Non-ASCII characters are treated specially. In the answer files, these characters are replaced by "?" character. No penalty or advantage will be given if no match will be made with those characters.

3. Pages to be Ignored

The following pages will not be used in the evaluation.

a) A page that does not contain the exact string of the name of a target person. For example, if the target name is "John Kennedy," and the name appears on a given page as "John F. Kennedy" only, the page will not be used.

b) A page that has two or more individuals sharing the same target name. For example, a page which contains "John Kennedy (Politician)" and "John Kennedy (Actor)" will not be used.

c) A page that displays search results from databases (e.g., DBLP and CiteSeer) or shopping sites (e.g., amazon.com and Yahoo! Shopping).

d) A page that is not written primarily in English.

e) A page on which the target name refers to a fictional character.

f) A page with fictional content, even if the target person in the fiction is a real-life figure.

4. Detailed definitions for each attribute

1. “Date of birth”

1a) An attribute value for “Date of birth” is the date when the target person was born. Even if a target person’s date of birth is expressed with only a year, month or day, it should be extracted as a value. Relative date expressions, such as “two years after Fred and Mary moved to England” should not be extracted.

2. “Birthplace”

2a) An attribute value for “Birthplace” is a location where the target person was born. It must be the name of a country, state, province, city, town, village or region. Non-names such as “manger” and “hospital”, or facility names such as “New York Hospital” cannot be extracted as values.

3. “Other name”

3a) An attribute value for “Other name” is any name of the target person other than the name indicated by the organizer. If a target person’s name does not appear exactly the same as the name provided for the search, it can be included as an attribute value for “Other name.” The values for this attribute include the expression of “surname, first name”, such as “Sekine, Satoshi” for “Satoshi Sekine”, as well as “JFK” or “John F. Kennedy” for “John Kennedy,” “Godzilla” for “Hideki Matsui,” or “The Godfather of Soul” for “James Brown”. Non-names such as “his wife” or “the president of XX company” should not be extracted.

4. “Occupation”

4a) An attribute value for “Occupation” is a name of an occupation of a target person. Verb phrases CANNOT be extracted as attribute values for “Occupation”. For example, the verb phrase “lectures Computer Science” cannot be extracted from the expression, “he lectures Computer Science.”

4b) “Occupation” can include a person’s current occupation, as well as any previous occupations.

4c) Names of specific entities, such as an affiliation, geographical and political entity (GPE), facility, or vehicle can not be as a part of a value for “Occupation”. However, other occupation names can be a part of a value for “Occupation”, like "Special Assistant to the President for Legislative Affairs," or “Parliamentary Secretary to the Minister for Employment, Education, Training and Youth Affairs.”

Expression	**Good**	**NG**
US Vice President	Vice President	US Vice President
Mayor of New York City	Mayor	Mayor of New York City

Development Director for NY	Development Director	Development Director for NY
Mid-Atlantic Manager	Manager	Mid-Atlantic Manager
Professor at MIT	Professor	Professor at MIT
Captain of PT-109	Captain	Captain of PT-109

4d) Common words of entities, such as an affiliation, GPE, facility, or vehicle can be as a part of a value for “Occupation”.

Expression	**Good**	**NG**
College Professor	College Professor	Professor
Taxi Driver	Taxi Driver	Driver
Software Developer	Software Developer	Developer

4e) An ordinal number expressing ranking is a part of a value for “Occupation” though an ordinal number expressing turn is not.

Expression	**Good**	**NG**
Second Infantry	Second Infantry	Infantry
35th President	President	35th President

4f) If it can be determined that the job of a target person is provisional or temporary (e.g., guest lecturer or conference organizer), it should not be extracted as a value of “Occupation.” (See “Ambiguous cases A.1.” below.)

5. <u>“Affiliation”</u>

5a) An attribute value for “Affiliation” is an organization name or a name of a group to which the target person belongs. The name of a department or study group can be extracted as an attribute value for “Affiliation”. For example, “Computer Science Department” or “Pattern Recognition and Image Processing Group” can be extracted as values.

5a) The name of an event CANNOT be extracted as a value for “Affiliation.” For example, “Tokyo International Film Festival Executive Committee” can be an affiliation, but “Tokyo International Film Festival” cannot.

5b) The system should extract current affiliations as well as any previous ones. However, the name of an alma mater should be extracted as an attribute value for “School”. If a target person was a student when the page was written, the name of his or her school should be considered a value for “Affiliation,” not “School”.

6. <u>“Work” (Note that this is not going to be evaluated at WePS2)</u>

6a) An attribute of “Work” is a name of any product created by the target person, such as a painting, piece of music, book, and so on. The name of an academic paper, journal article, speech and/or conference presentation cannot be extracted as an attribute value for “Work”. If they have been published as a book or issued as an electronic book, they can be extracted as values for “Work”.

6b) If a target person is an actor, the name of an episode of a TV drama in which he or she appears can be an attribute value for “Work”.

7. <u>“Award”</u>

7a) An attribute value for “award” is a name of an award the person has received.

8. “School”

8a) An attribute value for “School” is a name of an institution, including a kindergarten, elementary school, middle school and high school which a target person attended. A name of a department or a research center to which a target person belonged as a student cannot be values for “School.”

Expression	Good	NG
Sarada Ranganathan Department of Library Science, University of Madras, India	University of Madras	Sarada Ranganathan Department of Library Science, University of Madras

8b) If a target person is a student, the name of his or her school should be considered a value for “Affiliation,” not “School.”

9. “Major”

9a) An attribute value for “Major” is a name of an academic field in which a target person is specializing or specialized. Do not extract the name of a minor.

Expression	Good	NG
Associates degree in Early Childhood Education and a minor in Child Psychology	Early Childhood Education	Early Childhood Education
		Child Psychology

9b) Do not extract an academic field which is not clearly expressed as a target person’s major.

Expression	Good	NG
He studied mathematics.	N/A	mathematics

9c) If a part of an academic degree is the name of a major like the Master of Business Administration (MBA), do not extract the part as a value for "Major". The entire expression should be a “Degree”.

Expression	Good	NG
Master of Library Science	Major: N/A	Major: Library Science
	Degree: Master of Library Science	Degree: Master of Library Science

10. “Degree”

10a) An attribute value for “Degree” is a name of an academic degree a target person received. Do not extract very general expressions such as “postgraduate law degree,” or “advanced law degree” as values for “Degree”. Only the expressions which are explicitly mentioned that the target person received the degree should be extracted.

Expression	Good	NG
advanced law degree	Major: law	Major: law
	Degree: N/A	Degree: advanced

11. “Mentor”

11a) An attribute value for “mentor” is the name of any individual who is or has been a mentor to the target person. Mentors may include school teachers, sports coaches and/or advisors.

12. “Location” (Note that this is not going to be evaluated at WePS2)

12a) An attribute of “Location” is a name of a location where the target person was living at the time the page was written or the name of the location of his or her home or office.

12b) Attribute values for “Location” should not include the names of places visited by a target person or places where he or she used to live.

12c) If a target person is deceased, the name of the cemetery or other place where his or her remains are located can be considered an attribute value for “Location”.

12d) An attribute value for “Location” cannot be extracted from a value for “Affiliation”. For example, if an affiliation of a target person is “New York University,” you might assume his or her location to be “New York”. However, you should NOT use such information as an indicator for “Location”.

12e) The address of an affiliation or agent of a target person is considered an attribute value for “Location.” However, if a target person is a writer, the address of his or her publisher cannot be used as a value for “Location.” This rule also applies to **“Phone”, “Fax”, “Email”** and **“Web site”.**

13. “Nationality”

13a) An attribute value for “Nationality” is a country name or an adjective of nationality for where the target person has citizenship. It CANNOT be determined from a value for “Occupation.” For example, if a target person is “the President of the United States of America,” “United States of America” cannot be extracted as a value for “Nationality”.

14. “Relatives”

14a) An attribute value for “Relatives” is a name of a target person’s parents, siblings, children or former and current spouses. Other relatives including siblings-in-law, children-in-law and common-law spouses should not be extracted.

15. “Phone”

15a) An attribute value for “Phone” is a phone number used to reach the target person. It is not necessary to include international ID numbers or area codes if it is not expressed. An extension number can be extracted as an attribute value for “Phone.”

16. “Fax”

16a) An attribute value for “Fax” is a fax number used to reach the target person. It is not necessary to include international ID numbers or area codes if it is not expressed.

17. “Email”

17a) An attribute value for “Email” is any complete email address of the target person. Any link or unusable e-mail address such as those listed below are not extractable.

Mail to: *Allan Hanbury*
Email Andrew Powell
E-mail: Lastname AT cs DOT nyu DOT edu
sekine(here comes AT)cs.nyu.edu

18. “Web site”

18a) An attribute value for “Web site” is the URL of a Web page or weblog operated or authorized by a target person. The URL of the official site of an affiliation of a target person is considered a value for "Web site."

18b) Pages related to a target person, such as a page written on the books the person wrote or an unofficial fan site of the person CANNOT be used as a value. The URL of the official site of an event in which a target person is involved (e.g., a film festival or academic conference) CANNOT be extracted as a value.

18c) Values for URL need not include http:// if it is not expressed.

5. Ambiguous cases

Certain expressions can be ambiguous in some contexts. For example, “baseball player” can be extracted as a value for “Occupation” if a person is a professional baseball player. However, if it is mentioned that an individual plays baseball as a hobby, then “baseball” cannot be considered an occupation (See A1 for more detail in this case). The context surrounding a possible attribute value should be considered in order to determine the intentional meaning, and this will at times require background knowledge of real world topics. Examples include, but are not limited to, the following:

A1.“Occupation”, or not?

Some role names can refer to both occupations and non-professional roles. A role name can be an occupation, but it can also be the role of non-professional person. For example, if an individual is a professional writer, “author” can be extracted as a value for “Occupation”. However, if a person such as a university professor, whose occupation has already been identified, has written a book, ”author” would not be an extractable value for “Occupation”.

Expression: Author (Tony Abbott)

Occupation = Author: if Tony Abbott is a professional writer.
Occupation = Author: if Tony Abbott’s occupation is unknown, but found he wrote a book
Occupation = N/A: if Tony Abbott is a university professor and wrote a book
Occupation = N/A: if Tony Abbott wrote a scientific paper or just an essay for a weblog

Expression: He is a good baseball player.

Occupation = baseball player: if he is a professional baseball player
Occupation = N/A: if he plays baseball as a hobby

A2.“Affiliation” or “Location”?

A location name, such as a city, can be a part of a university name.

Expression: He has come back to Birmingham.

Location = Birmingham: if he has come back to Birmingham.
Affiliation = Birmingham: if he has come back to University of Birmingham.

A3. “Affiliation” or “Location”

A location name can be a part of a university name. The natural convention should be followed. For example, the University of California, Los Angeles is usually referred to as UCLA, but the University of Arizona, Tucson is not referred to as UAT.

Expression: He is at the University of California, Los Angeles.
Affiliation = University of California, Los Angeles
Location = N/A

Expression: He is at University of Arizona, Tucson.
Affiliation = University of Arizona
Location = Tucson

A4. “Occupation” or “Education”

The title, “Dr.” is an attribute value for “Education,” and if a target person is a medical doctor, the title can also be a value for “Occupation.”

Expression: Dr. Edward Fox

Occupation = Dr., Education = Dr.: if he has an MD
Occupation = N/A, Education = Dr.: if he has a PhD

6. Data Format

The data should be in the format of text files. A single file contains all information about a single person. The file name is the name of the parson (the white space should be replaced by underscore (_)) with “.txt” file extension (for example “Alexander_Macomb.txt”). Each line represents attribute information in a particular page, consisting ID number, attribute name and the list of values for the attribute separated by a tab (\t) character (See the example below and the training data). When you submit, all the files should be stored in the same directory. The directory should have the name of your submission ID or your institute name. You should make the all the data into a single tar, tgz or gip file when you send it to the organizer. Please follow how the training data is created.

```
8      Other name  ALEX. MACOMB.
8      Occupation  Brigadier General Gen.
9      Date of birth     4 February 1888   02/05/1888
9      Occupation  Captain     CPT
9      Affiliation United States Navy      USN
9      Location          Arlington National Cemetery
9      Relatives   Augustus Canfield Macomb      Ella Chelle McKelden
```

7. Evaluation

Evaluation will be conducted by comparing the system output and gold standard data created by annotators. The gold standard data will first be created by annotators before consulted the system output. Then, all the output over-generated by the systems may be checked by annotators to see if there are answers missing. Such a scheme has been chosen due to the success of the pooling scheme in the field of information retrieval.

The comparison will be done using recall, precision and F-measures for each individual attribute and for the overall answers.

Recall = (# of correctly identified attribute values by system) / (# of attribute values in gold data)

Precision = (# of correctly identified attribute values by system) / (# of attribute values the system produced)

F = 2 * Recall * Precision / (Recall + Precision)