/* default css */ table { font-size: 1em; line-height: inherit; border-collapse: collapse; } tr { text-align: left; } div, address, ol, ul, li, option, select { margin-top: 0px; margin-bottom: 0px; } p { margin: 0px; } pre { font-family: Courier New; white-space: pre-wrap; margin:0; } body { margin: 6px; padding: 0px; font-family: Verdana, sans-serif; font-size: 10pt; background-color: #ffffff; color: #000; } img { -moz-force-broken-image-icon: 1; } @media screen { html.pageview { background-color: #f3f3f3 !important; overflow-x: hidden; overflow-y: scroll; } body { min-height: 1100px; counter-reset: __goog_page__; } * html body { height: 1100px; } .pageview body { border-top: 1px solid #ccc; border-left: 1px solid #ccc; border-right: 2px solid #bbb; border-bottom: 2px solid #bbb; width: 648px !important; margin: 15px auto 25px; padding: 40px 50px; } /* IE6 */ * html { overflow-y: scroll; } * html.pageview body { overflow-x: auto; } /* Prevent repaint errors when scrolling in Safari. This "Star-7" css hack targets Safari 3.1, but not WebKit nightlies and presumably Safari 4. That's OK because this bug is fixed in WebKit nightlies/Safari 4 :-). */ html*#wys_frame::before { content: 'A0'; position: fixed; overflow: hidden; width: 0; height: 0; top: 0; left: 0; } .writely-callout-data { display: inline-block; width: 1px; height: 1px; overflow: hidden; margin-left: -1px; } .writely-footnote-marker { background-image: url('MISSING'); background-color: transparent; background-repeat: no-repeat; width: 7px; overflow: hidden; height: 16px; vertical-align: top; -moz-user-select: none; } .editor .writely-footnote-marker { cursor: move; } .writely-footnote-marker-highlight { background-position: -15px 0; -moz-user-select: text; } .writely-footnote-hide-selection ::-moz-selection, .writely-footnote-hide-selection::-moz-selection { background: transparent; } .writely-footnote-hide-selection ::selection, .writely-footnote-hide-selection::selection { background: transparent; } .writely-footnote-hide-selection { cursor: move; } /* Comments */ .writely-comment-yellow { background-color: #ffffd7; } .writely-comment-orange { background-color: #ffe3c0; } .writely-comment-pink { background-color: #ffd7ff; } .writely-comment-green { background-color: #d7ffd7; } .writely-comment-blue { background-color: #d7ffff; } .writely-comment-purple { background-color: #eed7ff; } .br_fix span+br:not(:-moz-last-node) { position:relative; left: -1ex } #cb-p-tgt { font-size: 8pt; padding: .4em; background-color: #ddd; color: #333; } #cb-p-tgt-can { text-decoration: underline; color: #36c; font-weight: bold; margin-left: 2em; } #cb-p-tgt .spin { width: 16px; height: 16px; background: url(//ssl.gstatic.com/docs/clipboard/spin_16o.gif) no-repeat; } } h6 { font-size: 8pt } h5 { font-size: 8pt } h4 { font-size: 10pt } h3 { font-size: 12pt } h2 { font-size: 14pt } h1 { font-size: 18pt } blockquote {padding: 10px; border: 1px #DDD dashed } .webkit-indent-blockquote { border: none; } a img {border: 0} .pb { border-width: 0; page-break-after: always; /* We don't want this to be resizeable, so enforce a width and height using !important */ height: 1px !important; width: 100% !important; } .editor .pb { border-top: 1px dashed #C0C0C0; border-bottom: 1px dashed #C0C0C0; } div.google_header, div.google_footer { position: relative; margin-top: 1em; margin-bottom: 1em; } /* Table of contents */ .editor div.writely-toc { background-color: #f3f3f3; border: 1px solid #ccc; } .writely-toc > ol { padding-left: 3em; font-weight: bold; } ol.writely-toc-subheading { padding-left: 1em; font-weight: normal; } /* IE6 only */ * html writely-toc ol { list-style-position: inside; } .writely-toc-none { list-style-type: none; } .writely-toc-decimal { list-style-type: decimal; }

.writely-toc-upper-alpha { list-style-type: upper-alpha; } .writely-toc-lower-alpha { list-style-type: lower-alpha; } .writely-toc-upper-roman { list-style-type: upper-roman; } .writely-toc-lower-roman { list-style-type: lower-roman; } .writely-toc-disc { list-style-type: disc; } /* Ordered lists converted to numbered lists can preserve ordered types, and vice versa. This is confusing, so disallow it */ ul[type="i"], ul[type="I"], ul[type="1"], ul[type="a"], ul[type="A"] { list-style-type: disc; } ol[type="disc"], ol[type="circle"], ol[type="square"] { list-style-type: decimal; } /* end default css */ /* default print css */ @media print { body { padding: 0; margin: 0; } div.google_header, div.google_footer { display: block; min-height: 0; border: none; } div.google_header { flow: static(header); } /* used to insert page numbers */ div.google_header::before, div.google_footer::before { position: absolute; top: 0; } div.google_footer { flow: static(footer); } /* always consider this element at the start of the doc */ div#google_footer { flow: static(footer, start); } span.google_pagenumber { content: counter(page); } span.google_pagecount { content: counter(pages); } .endnotes { page: endnote; } /* MLA specifies that endnotes title should be 1" margin from the top of the page. */ @page endnote { margin-top: 1in; } callout.google_footnote { display: prince-footnote; footnote-style-position: inside; /* These styles keep the footnote from taking on the style of the text surrounding the footnote marker. They can be overridden in the document CSS. */ color: #000; font-family: Verdana; font-size: 10.0pt; font-weight: normal; } /* Table of contents */ #WritelyTableOfContents a::after { content: leader('.') target-counter(attr(href), page); } #WritelyTableOfContents a { text-decoration: none; color: black; } /* Comments */ .writely-comment-yellow { background-color: #ffffd7; } .writely-comment-orange { background-color: #ffe3c0; } .writely-comment-pink { background-color: #ffd7ff; } .writely-comment-green { background-color: #d7ffd7; } .writely-comment-blue { background-color: #d7ffff; } .writely-comment-purple { background-color: #eed7ff; } } @page { @top { content: flow(header); } @bottom { content: flow(footer); } @footnotes { border-top: solid black thin; padding-top: 8pt; } } /* end default print css */ /* custom css */ /* end custom css */ /* ui edited css */ /* end ui edited css */ /* editor CSS */ .editor a:visited {color: #551A8B} .editor table.zeroBorder {border: 1px dotted gray} .editor table.zeroBorder td {border: 1px dotted gray} .editor table.zeroBorder th {border: 1px dotted gray} .editor div.google_header, .editor div.google_footer { border: 2px #DDDDDD dashed; position: static; width: 100%; min-height: 2em; } .editor .misspell {background-color: yellow} .editor .writely-comment { font-size: 9pt; line-height: 1.4; padding: 1px; border: 1px dashed #C0C0C0 } /* end editor CSS */

1. Overview

This subtask is to extract 16 kinds of “attribute values” of target individuals (i.e. cluster of Web pages)
. The organizers will distribute the target Web pages in their original format, (i.e., html), and the participants will be expected to cluster the documents according to the
different people sharing the name
(“clustering subtask”)
and extract certain biographical attributes for each person (
“attribute extraction subtask”
).

Note that in WePS-3 the attribute extraction subtask requires systems to participate in the document clustering

task

. In other words, u
nlike the WePS-2 AE task, attributes have to be assigned to each person profile
(e.g. cluster)
rather than to individual pages
. However,
systems are still required to specify the source of each attribute in their output.

All attributes to be extracted are listed in Table 1 below.

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

Award

Pulitzer Prize

7

School

Stanford University

8

Major

Mathematics

9

Degree

Ph.D.

10

Mentor

Tony Visconti

11

Nationality

American

12

Relatives

Jacqueline Bouvier

13

Phone

+1 (111) 111-1111

14

FAX

(111) 111-1111

15

Email

xxx@yyy.com

16

Web site

http://nlp.cs.nyu.edu

Table 1 Definition of 16 attributes of Person at WePS-2

In the following section, the general rules of the attribute extraction sub task will be explained. Section 3 provides participants with a detailed definition of each attribute as well as an explanation of potentially ambiguous cases. Section 4 explains the data format and 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 different 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
can
be extracted as one value. For example, the entire phrase “Tokyo, Japan”
can
be extracted from the expression, “He was born in Tokyo, Japan.”

c) However, for the same person, you are expected to extract only one mention of the duplicated attribute values. For example, if there are more than one document in a cluster, you should avoid to extract duplicated attribute values (e.g. several “Japan” for the Birthplace) from different pages. This is also the case if the mentions of the same value have variations (e.g. “New York University” and “NYU”, or “General” and “Gen.”). We will NOT give a penalty if a system produced duplicated values, but we will randomly choose only one value to the evaluation.

d) If a page contains a factual error, we will accept it as a correct attribute value . For example, both “1782” and “June 25, 1841” are correct 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.”

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

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 will be given either way.

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 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

3. 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.

1.

“**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.

1.

“**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.

1.

“**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.)

1.

“**Affiliation**”

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) It is OK to 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”.

1.

“**Award”**

6a) An attribute value for “award” is a name of an award the person has received.

1.

“**School”**

7a) 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

7b) If a target person is a student, the name of his or her school should be considered a value for “Affiliation,” not “School.”

1.

“**Major”**

8a) 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 Psychol

Early Childhood Education

Early Childhood Education

Child Psychology

8b) 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

8c) 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 $cience

Degree: Master of Library Science

Degree: Master of Library Science

1.

“**Degree”**

9a) 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

1.

“**Mentor”**

10a) 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.

1.

“**Nationality”**

11a) 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”.

1.

“**Relatives”**

12a) 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.

1.

“**Phone”**

13a) 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.”

1.

“**Fax**”

14a) 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.

1.

“Email”

15a) 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

1.

“**Web site”**

16a) 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."

16b) 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.

16c) Values for URL need not include http:// if it is not expressed.

4. 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

5. Data Format

Both the clustering and attribute extraction output must be provided in the same XML file. In this file each cluster of documents is specified by the element “entity”, which contains the list of grouped documents and the list of extracted attributes. For each attribute it's required to indicate

the type of attribute (date_of_birth, occupation, etc.), the source from which it was extracted (document ranking) and the value. The organizers will provide a detailed definition (DTD) of the XML output format and a validation script along with the WePS-3 trial data.

```
<clustering searchString="AMANDA LENTZ">

  <entity id="16" notes= "">

    <documents>

      <doc rank="17" notes= "" />

      <doc rank="66" notes= "" />
```

<doc rank="73" notes= "" />

<doc rank="51" notes= "from Huron" />

</documents>

<attributes>

<attr

type="date_of_birth" source="17" notes= "">4th August 1979</attr>

<attr type="occupation" source="17" notes= "" >Painter</attr>

</attributes>

</entity>

[...]

</clustering>

6. Evaluation

Attribute extraction will be done on the clusters of the selected two people, not all the people of the name (or all clusters of the name).

Participating systems will be

evaluated based on the attributes they attach to the cluster which has the best F-measure (with the weight of precision to recall set to 2) in the clustering task. So, the s
ystems are required

to extract values for each attribute for all clusters

.

The systems are requested to report

the document ID from which they extracted each attribute value.

The attribute extraction task

evaluation will be done by a pool of the system outputs, so coverage is not guaranteed on the attribute annotations.