Extracting the Author of Web Pages

Yoshikiyo Kato, Daisuke Kawahara, Kentaro Inui (NICT)
Sadao Kurohashi (NICT / Kyoto Univ.)
Tomohide Shibata (Kyoto Univ.)
Information Credibility and the Source of Information

• The source of information (information sender) is one of the important elements for judging the credibility of information
  – Anonymous vs. a person with a certain background
  – Business selling health food products vs. medical expert

• Elements of the credibility of information sender
  – Competence: expertise, in a position to know
  – Intention: motivation for disseminating credible information or misinformation

• Before we can analyze these characteristics, we need to identify the information sender
  $\rightarrow$ Focus of this study
Who is the information sender of this page?

NICT (National Institute of Information and Communications Technology)

Which one is the information sender of this page?

Both!

However, their roles (and the responsibilities concomitant with them) are not the same.
Identifying Information Sender Configuration

• Goal:
  – To identify the information senders and their roles in publishing the information as information sender configuration

• Elements of Information Sender Configuration
  – Information Sender (e.g. “NICT”)
  – Sender Class (e.g. Government)
  – Configuration Type (e.g. bunch)

• Example
  (bunch,
   (Government, “NICT”),
   (-, “Miyahara Hideo”, “President”, -))
Sender Class

• Information senders are categorized into sender classes

• Axes of classification:
  • Individual vs. Organization
  • Profit vs. Nonprofit Orgs.
  • Expertise (Univ., Medical, etc.)
  • Social function (Press)
  • Anonymous vs. Real Name

1. Organization
   (a) Profit Organization
      i. Company
      ii. Industry Group
   (b) Nonprofit Organization
      i. Academic Society
      ii. Government
      iii. Political Organization
      iv. Public Service Corporation, Nonprofit Organization
      v. University
      vi. Voluntary Association
      vii. Education Institution
      viii. Medical Institution
   (c) Press
      i. Broadcasting Station
      ii. Newspaper
      iii. Publisher

2. Individual
   (a) Real Name
   (b) Anonymous, Screen Name
Configuration Type

- Configuration type indicates the relationship between information senders
- Six types have been defined
  - The difference among types are concerned with the attribution of responsibility for the information published

<table>
<thead>
<tr>
<th>Configuration Type</th>
<th>Information Sender</th>
<th>Site Operator</th>
</tr>
</thead>
<tbody>
<tr>
<td>singleton</td>
<td>&lt;belongs to&gt;</td>
<td></td>
</tr>
<tr>
<td>bunch</td>
<td>&lt;publishes&gt;</td>
<td></td>
</tr>
<tr>
<td>publish</td>
<td>&lt;quotes&gt;</td>
<td></td>
</tr>
<tr>
<td>quote</td>
<td>&lt;uses&gt;</td>
<td></td>
</tr>
<tr>
<td>service</td>
<td></td>
<td>&lt;belongs to&gt;</td>
</tr>
<tr>
<td>composite</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
How ISCs are Used?

Show distribution of opinions according to sender class
(A kind of poll on the Web)
Toward Identifying Information Sender Configuration (ISC)

• Undertakings so far: tackling sub-problems of ISC identification
  1. Site operator identification
     (e.g. identifying “NICT” as the site operator)
        Precision: 70%
        (N.B.: achieves 80% when blogs/forums excluded)
  2. Sender classification
     (e.g. classifying “NICT” as Government)
        Precision: 74%
  3. Author identification
     (i.e. identifying “Hideo Miyahara” as the author)
        Precision: 59%
Process of Author Extraction

Input
- URL
- HTML file

Sentence Extraction

Sentences

Sentence Extraction

Author Candidates
- A
- B
- C
- D
- E

Compound Noun Extraction

Rank Candidates

Output
1. C
2. A
3. E
4. D
5. B
Sentence Extraction

• Extract text from HTML
• Segment texts into sentences
• Apply Japanese parser (KNP) to the sentences
• Retain sentences that satisfy the following conditions (to filter out normal sentences):
  – The ratio of the number particles other than ‘no’ (“of”) to the number of morphemes in the sentence is less than a certain threshold
  – Does not contain compound particles with verb (e.g. ‘~ ni tsuite’ (“concerning ~”))
Compound Noun Extraction

• Extract clauses that satisfy the following conditions:
  – At least one named entity typed either as a person or an organization is recognized (using NER functionality of KNP)
  – The sentence contains one or more morphemes typed as not-in-dictionary word
  – The last morpheme but particles has a POS tag of person name suffix or organization name suffix

• Extract compound noun from qualifying clauses
Example of author extraction

URL: http://www.nict.go.jp/about/message.html
Title: NICT 独立行政法人 情報通信研究機構

Author Candidates
Ranking Author Candidates

• Rank author candidates according to how likely they are indeed the author of the page
• Model: Ranking SVM (Joachims 2002)
• Labels: scores based on the match of the candidate to the human annotation
  – Exact match: 2
  – Partial match: 1
  – No match: 0
• Features
  – TF
  – POS tags (especially, # of PERSON, LOCATION, ORGANIZATION words)
  – Tokens (separate features for the starting and ending tokens)
  – HTML tags
  – Context words
  – Distance from the main content
Hypothesis: Author names appear in the proximity of the main content of the page.
Recognition of Main Content based on Text Volume

\[ r(n) = \frac{v(n)}{v(root)} \]

\[ \arg\min_n r(n) \text{ such that } r(n) \geq r_{th} \]

\[ v(n) \text{ The volume of text that the sub-tree rooted at node } n \text{ contains} \]
Distance of Elements in a Web Page

- *Rendered Distance*: Distance based on positions of elements after rendering
  - Pros:
    - Reflects the distance that users actually see
  - Cons
    - Rendering is computationally expensive
    - Needs style sheets and image files for accurate rendering

- *Document Structure Distance*: Distance based on document structure
  - Pros:
    - Computationally cheaper than rendering
    - Does not require style sheet or image files
  - Cons:
    - Does not necessarily reflect the distance that the user actually sees
Flattening of DOM Tree

• Purpose
  – Want to measure distance of elements in DOM
  – Want it to be as close to the rendered distance

• A straightforward measure
  – Path length between the nodes in the DOM tree
  – Problem:
    DOM trees tend to have deeply nested structure and
does not necessarily reflect the rendered distance

• Flattening of DOM
  – Convert a DOM tree into a sequence of block-level
  elements that directly has text nodes as its children
Example of DOM flattening
Distance from the Main Content

$d_m$ Distance from the main content measured on the flattened DOM (0 when the element is inside the main content)

$d_b$ Distance from the boundary of the main content measured on the flattened DOM

$d_c$ Distance from the center of the main content measured on the flattened DOM normalized by the size of the main content
Distance from the Main Content

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th>1</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>p</td>
<td>NICT 独立行政法人情報通信研究機構</td>
<td>17</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>span</td>
<td>ナビゲーションをスキップして本文へ</td>
<td>16</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>form</td>
<td>サイト内検索</td>
<td>15</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>li</td>
<td>ナビゲーション一覧を開く</td>
<td>14</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>li</td>
<td>English</td>
<td>13</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>p</td>
<td>現在位置: ホームの中の機構案内の…</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>h1</td>
<td>理事長あいさつ</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>div</td>
<td>理事長 宮原秀夫</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>p</td>
<td>情報通信研究機構は、来るべきユビキタネット社会を支える情報通信技術の…</td>
<td>0</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
Experiment

• Dataset
  – Information Credibility Evaluation dataset [Miyamori et al., 2008]
    • ~2000 pages (20 topics * 100 pages)
    • Each page has ISC annotated by human annotator
  – Used only those pages that has information sender type of bunch or publish (~400 pages)

• Evaluation
  – Topic-wise 20-fold cross-validation
  – Evaluation measure:
    Precision of ranking at cut off rank of 1, 3, and 5

• Results
  – Precision: 58.6% (@1), 72.0 (@3), 75.2 (@5)
    (If all the candidates considered: 84.7%)
Successful Example (1) : Government Notice

Kitajima, General Manager, Office of Health Policy on Newly Developed Food

1. 厚生労働省では、アガリクスを含む製品が広く流通していることから、アガリクスを含む製品によることが明らかに健康被害は報告されていませんが、肝障害の恐れ等複数の手例の件で補償対象に挙げられていること等から、厚生労働省食品衛生研究所において、アガリクスを含む市販の食品等の発生実験を実施してきました。この結果、現在実施しているラットを用いた中調度投与発症試験（第1）において、1製品に発がんプロモーション作用（※2）が認められたとの中間報告を受けました（表1参照）。

2. これを踏まえ、本日、厚生労働省は食品安全委員会に、これらのアガリクスを含む製品の食品健康新規影響評価を依頼することを発表しました。今回、3製品のうちの1製品については、食品としての販売を暫定的に禁止することについて、他の2製品については製品の安全性についての調査を依頼することとしました。

3. 今回の中間結果は、ラットに投与発がんプロモーション作用を認めたものであり、ヒトに対してどうかがんを引き起こすような結果ではありませんが、ヒトへの健康被害を未然に防ぐため、厚生労働省では、次のような対応を講じることとしています。

(1)アガリクスの確認と分析
(2)調査対象掛差を検証するため、当該製品の取扱を検証する（※3）こととして、アガリクスに関するQ&Aを厚生労働省のホームページに掲載し、適切な情報提供
(3)食品調査等、関係機関等に対し、食品の協力提供するための通達を発出
(4)厚生労働省にアガリクスを含む製品に関する相談専用電話を設置

4. なお、他の2製品は、1製品（仙台製薬「ゴールド」販売者：株式会社サファリ（※4））については、遺伝性毒性試験は陰性で、発がんプロモーション試験が実施中であり、遺伝性発病の増加は確認されていません（※5）。他の1製品（アガリクス製品「ABC」販売者：株式会社サファリ）については、遺伝性が陰性で、ラットにおける発がんプロモーション試験結果で現在遺伝性発病の増加は認められていないとの報告を受け、結果概要は出しておりません。これら2製品については、調査結果が出次第、食品安全委員会にその内容を報告するとともに公表することとしています。
 успешный пример (2) : статья из газеты

Кумико Накаима, отдел медицинской информации
Successful Example (3) : Article from a Web site

System Output: 秋沢

今日こそは“ダイエットを成功させる！”という方へ

今年こそは“ダイエットを成功させる！”という方へ

年末年始はたくさん食べると、運動量が少なく、「最近、気がうっとうしいよ」とお悩みの方はきっと少なくないはずです。

そこで、「体重をどうにか減らしたい！」とダイエット計画を立てている方に参考にしていただきたいのが、以下に示すサプリメントです！

1月1日から3月31日までの間、 Depository Library に展示されます。
Failed Example (1): The author is a newswire (not a person)

System Output: 平松

News section of a portal site

Answer: J-CASTニュース
Failed Example (2) : More than one person name near the main content

Akira Misaki, Professor Emeritus, Osaka City Univ.

Answer: 木村正人 (Masato Kimura)
漢方で1日1善
あなたにも当てはまるケースはありませんか？
当店に来られる漢方薬をの中から症例を漢方薬処方でできるだけ毎日記録していくようにいたします。
ずっと一人で悩んでいたこと、知ったかったこと…ここで見つけてください。

Failed Example (3) : Blog

System Output: 鈴木佳代

Answer: 古村学
Learned Model: Positive Features

<table>
<thead>
<tr>
<th>Summed weights</th>
<th>Features</th>
<th>Features based on linguistic analysis</th>
<th>Features based on document structure</th>
<th>Context of the author name</th>
<th>Tokens that are specific to certain instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>486.975525</td>
<td>person</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>343.295375</td>
<td>context_tag_font</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>314.868000</td>
<td>num_morph</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>303.745805</td>
<td>context_tag_select</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>303.745805</td>
<td>context_tag_option</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>270.825498</td>
<td>context_tag_b</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>154.405898</td>
<td>WORD_IPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>137.443534</td>
<td>location</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>131.968790</td>
<td>startword_IPS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>124.140339</td>
<td>context_word_日</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>118.839416</td>
<td>startword_松崎</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>118.839416</td>
<td>WORD_松崎</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Learned Model: Negative Features

<table>
<thead>
<tr>
<th>Summed weights</th>
<th>Features</th>
<th>Features based on document structure</th>
</tr>
</thead>
<tbody>
<tr>
<td>-189.670375</td>
<td>context_tag_span</td>
<td></td>
</tr>
<tr>
<td>-132.047854</td>
<td>context_tag_img</td>
<td></td>
</tr>
<tr>
<td>-104.351307</td>
<td>distance_from_main_edge</td>
<td>Distance based feature</td>
</tr>
<tr>
<td>-38.114124</td>
<td>tag_meta_keywords</td>
<td></td>
</tr>
<tr>
<td>-38.060345</td>
<td>context_word_方</td>
<td>Relatively general context words</td>
</tr>
<tr>
<td>-33.739126</td>
<td>WORD_ダイエット</td>
<td></td>
</tr>
<tr>
<td>-31.354978</td>
<td>WORD_会</td>
<td>Topic specific tokens</td>
</tr>
<tr>
<td>-30.968525</td>
<td>context_word_写真</td>
<td>Relatively general tokens</td>
</tr>
<tr>
<td>-29.672442</td>
<td>startword_ダイエット</td>
<td></td>
</tr>
<tr>
<td>-28.682960</td>
<td>startword_日本</td>
<td></td>
</tr>
<tr>
<td>-28.491197</td>
<td>context_word_サイト</td>
<td></td>
</tr>
<tr>
<td>-27.058144</td>
<td>WORD_市</td>
<td></td>
</tr>
</tbody>
</table>

Weights of other distance features:  
-0.673151  distance_from_main  
-19.485953  distance_from_main_center
Observations

• The method works when the author name appears near the main content (*this is what we have expected*)
• Performs poorly when:
  1. the author name do not appear near the main content (as in blogs)
  2. there are more than one person names near the main content
  3. the author is not a person but an organization
Possible Improvements

• Add more features:
  – which can be used to identify author names that are relatively far from the main content
  – which can distinguish the author name from other person names which are not the author

• More data:
  – There may be not enough instances where the author is an organization for the model to learn
    30 / 400 ~ 7.5% (rest are individuals)

• Web page structure recognition
  – Recognize the semantics of the parts of a Web page
    • Main article, profile, biography, link list, ads, etc.
Summary

• Approach to information credibility analysis from the aspect of *information sender*
• Description of information senders and their relationships in terms of *information sender configuration (ISC)*
• Extraction of the author of Web pages as a sub-problem of ISC identification
• Evaluation results: Precision at ~60% (75% @3)
• Challenges ahead:
  – Improving precision
  – Dealing with multiple authors
    (forum posts, product review comments)
  – Identifying parts of a Web page that come from different authors, and their corresponding authors
  – Expertise analysis
Merci beaucoup

ykato@nict.go.jp