Automatic Analysis of Author Judgment in Scientific Articles Based on Semantic Annotation

Marc Bertin, Iana Atanassova and Jean-Pierre Desclés

Paris-Sorbonne University, LaLIC Laboratory

20 May 2009
FLAIRS 2009
Outline

1 Introduction
   - Problem
   - Semantic Annotation Strategy

2 Implementation
   - Text processing
   - Corpus
   - Semantic Map

3 Demonstration

4 Evaluation
   - Methodology
   - Results

5 Conclusion
1 Introduction
   • Problem
   • Semantic Annotation Strategy

2 Implementation
   • Text processing
   • Corpus
   • Semantic Map

3 Demonstration

4 Evaluation
   • Methodology
   • Results

5 Conclusion
1. How can we use the bibliographic citations of authors in texts?

2. Our aim is to identify relations between authors and more precisely the nature of these relations by an automatic semantic annotation of citations.

3. There are various motivations to cite and there are many functions of citation. Citation is a complex phenomenon.

4. Citation analysis requires linguistic approaches for the categorization of the relations between authors.
How can we use the bibliographic citations of authors in texts?

Our aim is to identify relations between authors and more precisely the nature of these relations by an automatic semantic annotation of citations.

There are various motivations to cite and there are many functions of citation. Citation is a complex phenomenon.

Citation analysis requires linguistic approaches for the categorization of the relations between authors.
1. How can we use the bibliographic citations of authors in texts?
2. Our aim is to identify relations between authors and more precisely the nature of these relations by an automatic semantic annotation of citations.
3. There are various motivations to cite and there are many functions of citation. Citation is a complex phenomenon.
4. Citation analysis requires linguistic approaches for the categorization of the relations between authors.
1. How can we use the bibliographic citations of authors in texts?
2. Our aim is to identify relations between authors and more precisely the nature of these relations by an automatic semantic annotation of citations.
3. There are various motivations to cite and there are many functions of citation. Citation is a complex phenomenon.
4. Citation analysis requires linguistic approaches for the categorization of the relations between authors.
Localization of relevant text segments:
- Segmentation into sentences
- Identification of indexed references in the sentences: by finite state automata

Localization of the textual segments in which we will most probably find the judgment of the author on another author:
- Hypothesis: author judgments are localized in the textual space close to an indexed reference. Segments containing references carry potentially information on the type of relation between the authors.

Automatic semantic annotation: Contextual Exploration Method (see Desclés 2006)

Exploitation of the annotation and text navigation
1. Localization of relevant text segments:
   - Segmentation into sentences
   - Identification of indexed references in the sentences: by finite state automata

2. Localization of the textual segments in which we will most probably find the judgment of the author on another author:
   - Hypothesis: author judgments are localized in the textual space close to an indexed reference. Segments containing references carry potentially information on the type of relation between the authors.

3. Automatic semantic annotation: Contextual Exploration Method (see Desclés 2006)

4. Exploitation of the annotation and text navigation
1 Localization of relevant text segments:
   - Segmentation into sentences
   - Identification of indexed references in the sentences: by finite state automata

2 Localization of the textual segments in which we will most probably find the judgment of the author on another author:
   - Hypothesis: author judgments are localized in the textual space close to an indexed reference. Segments containing references carry potentially information on the type of relation between the authors.

3 Automatic semantic annotation: Contextual Exploration Method (see Desclés 2006)

4 Exploitation of the annotation and text navigation
1. Localization of relevant text segments:
   - Segmentation into sentences
   - Identification of indexed references in the sentences: by finite state automata

2. Localization of the textual segments in which we will most probably find the judgment of the author on another author:
   - Hypothesis: author judgments are localized in the textual space close to an indexed reference. Segments containing references carry potentially information on the type of relation between the authors.

3. Automatic semantic annotation: Contextual Exploration Method (see Desclés 2006)

4. Exploitation of the annotation and text navigation
Localization of relevant text segments:

1. Segmentation into sentences
2. Identification of indexed references in the sentences: by finite state automata

Localization of the textual segments in which we will most probably find the judgment of the author on another author:

- Hypothesis: author judgments are localized in the textual space close to an indexed reference. Segments containing references carry potentially information on the type of relation between the authors.

Automatic semantic annotation: Contextual Exploration Method (see Desclés 2006)

Exploitation of the annotation and text navigation
Introduction

1. Problem
2. Semantic Annotation Strategy

Implementation

3. Text processing
4. Corpus
5. Semantic Map

Demonstration

Evaluation

1. Methodology
2. Results

Conclusion
Linguistic approach: Contextual Exploration Method

Semantic annotation tool: the EXCOM (Multilingual Contextual Exploration) system, developed by the LaLIC Laboratory

The major objective for the EXCOM system is to explore the semantics of texts for enhancing information extraction and retrieval through automatic annotation of semantic relations.
According to our protocol, we take into consideration texts containing indexed references and bibliography, such as scientific articles, reports, PHD dissertations, publications, etc.

Our corpora are in French, in the domains of Social Sciences and Humanities.
According to our protocol, we take into consideration texts containing indexed references and bibliography, such as scientific articles, reports, PHD dissertations, publications, etc.

Our corpora are in French, in the domains of Social Sciences and Humanities.
<table>
<thead>
<tr>
<th>Corpus</th>
<th>Language</th>
<th>Coverage</th>
<th>Format</th>
</tr>
</thead>
<tbody>
<tr>
<td>CALS</td>
<td>fr</td>
<td>33 texts</td>
<td>pdf</td>
</tr>
<tr>
<td>LaLIC</td>
<td>fr</td>
<td>8 texts</td>
<td>doc/pdf</td>
</tr>
<tr>
<td>TALN</td>
<td>fr/eng</td>
<td>1999-2005</td>
<td>pdf</td>
</tr>
<tr>
<td>Intellectica</td>
<td>fr/eng</td>
<td>1991-2002</td>
<td>pdf</td>
</tr>
<tr>
<td>IRISA</td>
<td>fr/eng</td>
<td>1984-2006</td>
<td>pdf/ps</td>
</tr>
<tr>
<td>PhD Theses</td>
<td>fr</td>
<td>6 PhD theses</td>
<td>pdf</td>
</tr>
</tbody>
</table>
Result:

Observations on individual MTs in a microscopic flow cell (Walker et al 1991) showed that polymerizing MTs transit very rapidly (within 14 s) upon dilution, suggesting that the cap size is fairly small, less than 100 subunits.

Measurements of the Pos of PBM vesicles by Niemietz & Tyerman (2000) yielded values that were lower than those measured by Rivers et al. (1997).

Method:

Extraction of DNA from filter samples followed a modification of a method employed by Fuhrman et al. [33] as described by Abell and Bowman [20].

This model was based on early observations of a relatively long kinetic lag between tubulin polymerization and GTP hydrolysis (Carlier & Pantaloni 1981).
Result:

"Observations on individual MTs in a microscopic flow cell (Walker et al 1991) showed that polymerizing MTs transit very rapidly (within 14 s) upon dilution, suggesting that the cap size is fairly small, less than 100 subunits."
Result:

"Observations on individual MTs in a microscopic flow cell (Walker et al 1991) showed that polymerizing MTs transit very rapidly (within 14 s) upon dilution, suggesting that the cap size is fairly small, less than 100 subunits."

"Measurements of the Pos of PBM vesicles by Niemietz & Tyerman (2000) yielded values that were lower than those measured by Rivers et al. (1997)."
Result:

"Observations on individual MTs in a microscopic flow cell (Walker et al. 1991) showed that polymerizing MTs transit very rapidly (within 14 s) upon dilution, suggesting that the cap size is fairly small, less than 100 subunits."

"Measurements of the Pos of PBM vesicles by Niemietz & Tyerman (2000) yielded values that were lower than those measured by Rivers et al. (1997)."

Method:
Result:

"Observations on individual MTs in a microscopic flow cell (Walker et al 1991) showed that polymerizing MTs transit very rapidly (within 14 s) upon dilution, suggesting that the cap size is fairly small, less than 100 subunits."

"Measurements of the Pos of PBM vesicles by Niemietz & Tyerman (2000) yielded values that were lower than those measured by Rivers et al. (1997)."

Method:

"Extraction of DNA from filter samples followed a modification of a method employed by Fuhrman et al. [33] as described by Abell and Bowman [20]."
Result:

"Observations on individual MTs in a microscopic flow cell (Walker et al 1991) showed that polymerizing MTs transit very rapidly (within 14 s) upon dilution, suggesting that the cap size is fairly small, less than 100 subunits."

"Measurements of the Pos of PBM vesicles by Niemietz & Tyerman (2000) yielded values that were lower than those measured by Rivers et al. (1997)."

Method:

"Extraction of DNA from filter samples followed a modification of a method employed by Fuhrman et al. [33] as described by Abell and Bowman [20]."

"This model was based on early observations of a relatively long kinetic lag between tubulin polymerization and GTP hydrolysis (Carlier & Pantaloni 1981)."
Information:

- It was surmised that this was due to bacterial cells dispersing from particles as the particles decompose and sink, a phenomenon originally proposed by Azam [46].
- Samuels et al. (35,36) reported similar results with the powdery mildew cucumber pathosystem, suggesting that in-soluble Si deposition is a common phenomenon in both dicots and monocots.
- This structure was originally postulated to be a cap of GTP-tubulin (Mitchison & Kirschner 1984a).
Information:

"It was surmised that this was due to bacterial cells dispersing from particles as the particles decompose and sink, a phenomenon originally proposed by Azam [46]."
Information:

"It was surmised that this was due to bacterial cells dispersing from particles as the particles decompose and sink, a phenomenon originally proposed by Azam [46]."

Similarity:
Semantic annotation: examples (2)

- Information:
  "It was surmised that this was due to bacterial cells dispersing from particles as the particles decompose and sink, a phenomenon originally proposed by Azam [46]."

- Similarity:
  "Samuels et al. (35,36) reported similar results with the powdery mildewcucumber pathosystem, suggesting that in-soluble Si deposition is a common phenomenon in both dicots and monocots."
Information:

"It was surmised that this was due to bacterial cells dispersing from particles as the particles decompose and sink, a phenomenon originally proposed by Azam [46]."

Similarity:

"Samuels et al. (35,36) reported similar results with the powdery mildewcucumber pathosystem, suggesting that in-soluble Si deposition is a common phenomenon in both dicots and monocots."

Hypothesis:
Information:

"It was surmised that this was due to bacterial cells dispersing from particles as the particles decompose and sink, a phenomenon originally proposed by Azam [46]."

Similarity:

"Samuels et al. (35,36) reported similar results with the powdery mildewcucumber pathosystem, suggesting that in-soluble Si deposition is a common phenomenon in both dicots and monocots."

Hypothesis:

"This structure was originally postulated to be a cap of GTP-tubulin (Mitchison & Kirschner 1984a)."
1 Introduction
   - Problem
   - Semantic Annotation Strategy

2 Implementation
   - Text processing
   - Corpus
   - Semantic Map

3 Demonstration

4 Evaluation
   - Methodology
   - Results

5 Conclusion
• Information retrieval
• Bibliosemantics
• Categorization: PhD theses
Introduction
- Problem
- Semantic Annotation Strategy

Implementation
- Text processing
- Corpus
- Semantic Map

Demonstration

Evaluation
- Methodology
- Results

Conclusion
Evaluations

- First evaluation: measuring the accuracy of the retained indicators, or indexed references, which have been identified automatically by the Finite State Automata.
Evaluations

- First evaluation: measuring the accuracy of the retained indicators, or indexed references, which have been identified automatically by the Finite State Automata.

- Second evaluation: carrying out a session of concordance between human judges in order to evaluate the rates of agreement between them by the Kappa coefficient.
Evaluation 1: Precision and Recall measures

Measuring the capacity of the system to correctly identify the textual segments containing indicators:

- results, by taking into consideration only the indexed references in texts:

  \[
  \begin{array}{c|c}
  \text{Recall} & \text{Precision} \\
  \hline
  91,09\% & 98,91\% \\
  \end{array}
  \]

- results, by taking into consideration also the named entities in the corpus:

  \[
  \begin{array}{c|c}
  \text{Recall} & \text{Precision} \\
  \hline
  67,15\% & 98,91\% \\
  \end{array}
  \]
<table>
<thead>
<tr>
<th>Nature</th>
<th>Indexed reference</th>
<th>Named entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identification</td>
<td>Regular expression</td>
<td>Named entity identification</td>
</tr>
<tr>
<td>Norms</td>
<td>ISO 690 and ISO 690-2</td>
<td></td>
</tr>
<tr>
<td>Epistemology</td>
<td>Frontier knowledge</td>
<td>Core knowledge</td>
</tr>
<tr>
<td>Out of context comprehension</td>
<td>None</td>
<td>Researcher from the domain</td>
</tr>
</tbody>
</table>

Growing complexity for the identification
Cohen’s weighed Kappa coefficient (Cohen 1960) provides a method to measure numerically the agreement between two or more observers or methods in the case when the judgments are qualitative in nature.

<table>
<thead>
<tr>
<th>Judge B</th>
<th>Answers</th>
<th>Correct</th>
<th>Incorrect</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>77</td>
<td>10</td>
<td>87</td>
<td></td>
</tr>
<tr>
<td>Incorrect</td>
<td>6</td>
<td>7</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>83</td>
<td>17</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

\[ K = 0.83 \]
1 Introduction
   • Problem
   • Semantic Annotation Strategy

2 Implementation
   • Text processing
   • Corpus
   • Semantic Map

3 Demonstration

4 Evaluation
   • Methodology
   • Results

5 Conclusion
We have already developed:

- Categorization of relations between authors using semantic annotation
- Analysis of the functions of citation
- Categorization of publications or sets of publications
We have already developed:

- Relevant information extraction
- Concept identification
- Categorized text syntheses
- Text navigation

Marc Bertin, Iana Atanassova and Jean-Pierre Desclés

Automatic Analysis of Author Judgment
What can we do next?

- Science policy by establishing guidelines and setting priorities
- Detecting *emergence* and *innovation*
- New method for mapping science
- Establishing author networks
Thank you for your attention!

Further information:
marc.bertin@paris-sorbonne.fr
iana.atanassova@gmail.com
jean-pierre.descles@paris-sorbonne.fr

Bibliography


Moed, H. 2005. Citation Analysis in Research Evaluation. Springer.

Small, H. 1982. Citation context analysis. B. Dervin and M. Voigt (Eds.), Progress in communication sciences 3:287310.