Scientific Paper Analysis: Knowledge Discovery through Structural Document Understanding

Yuji Matsumoto
Nara Institute of Science and Technology
Nara, Japan
Objective of the Project

- Developing and deepening Text/Document analysis technologies
- Implementing similarity-based document retrieval and recommendation
- Enabling knowledge acquisition, summarization, visualization, survey
- All of those will be done through various levels of structural similarity analysis of documents
Overview of the Project

Text/Document Analysis (T1, T5)

Knowledge Base Fertilization (T2, T3, T4)

Structural Similarity Analysis (T6, T7)

User Interface / Survey Generation (T8, T9)

Inner-Document Analysis
- Text/Document Analysis
- Concept/Relation Analysis
- Predicate-argument str analysis
- Event Chain Analysis
- Document Structure Analysis
- Document Summarization

Inter-Document Analysis
- Citation Analysis
- Document Relation
- Document Similarity
- Multi-Document Summarization
- Survey Generation

User Interface / Document Visualization
Research Groups

- **G0: Matsumoto, Shimbo, Shindo, Noji, ... (NAIST)**
  - Semantic and structure analysis of Bio-/Material science documents
  - Collaboration with researchers in Biology and Material Sciences

- **G1: Satoh (NII), Nguyen (JAIST)**
  - Similar document retrieval for judicial precedents

- **G2: Inui, Inoue (Tohoku U)**
  - Development of argumentation grammar and annotation study

- **G3: Aizawa, Miyao, Abekawa(NII), Nanba(Hiroshima City U)**
  - Document structure analysis

- **G4: Tsuruoka (U Tokyo)**
  - Text Summarizer of Academic Papers

- **G5: Mori (U Tokyo)**
  - Citation Analysis: Detecting Research Trend of Academic Fields

- **G6: Kano (Shizuoka U)**
  - Table Data Extraction from Neuroscience Papers
Method 1: Document/Text Analysis

Sentence analysis
- Tokenization
- POS tagging
- Sentence boundary detection
- Named Entity / Term chunking
- Parsing
  - Multi-word expressions
  - Coordination structure analysis
  - Complex sentence structure analysis

Document Analysis
- Document structure analysis
- Document similarity
- Summarization
Method 2: Structural Similarity Analysis

Structural Similarity Analysis at various levels
- Document structure
- Citation
- Relation between entities
- Event
- Event chain
- Domain-dependent research structure
Method 3: Similarity-based Knowledge Extraction

- Knowledge Extraction from scientific papers
- Relation extraction: Database fertilization
  - Biological Science
    - KNApSAcK Database: Metabolite-Plant Species Database
    - KEGG (Kyoto Encyclopedia of Genes and Genomes) database
  - Material Science
    - Extraction of Thermoelectric elements and their properties
- Data Annotation
  - Annotation Tool for PDF documents
Text Analysis

Sentence analysis
- Tokenization
- POS tagging
- Sentence boundary detection
- Named Entity / Term chunking
- Parsing
  - Multi-word expressions
  - Coordination structure analysis
  - Complex sentence structure analysis
NeuralNet-based POS tagger

POS tagging

Word-level CNN

Character-level CNN

IN  VBG  PRP  NN  RP

in  getting  their  money  back

i n  g e t t i n g  b a c k
Typical Coordinate Structure appearing in Scientific Papers

Identification of coordinate structure helps improve parsing accuracy

- “Median times to progression and median survival times were 6.1 months and 8.9 months in arm A and 7.2 months and 9.5 months in arm B.”

↓

- “Median times ... were 6.1 months in arm A”
- “Median times ... were 7.2 months in arm B”
- “median survival times were 8.9 months in arm A”
- “median survival times were 9.5 months in arm B”
Dependency and Coordinate Structure Annotation

ChaKi ©NAIST
Other Sentence Analysis Tasks

- Complex Sentence Structure Analysis
- Multi-Word Expressions
  - Lexicons of English complex sentence structures and multi-word expressions
Method 2: Structural Similarity Analysis

Structural Similarity Analysis at various levels

- Document structure
- Citation
- Relation between entities
- Event
- Event chain

- Domain-dependent research structure
Similarity at Document Level

**Generation of Structured Abstract from scientific papers**
- Structured Abstract: Consisting of Background, Objective, Method, Result, Conclusion
- Defining similarity/dis-similarity between documents at Objective/Method/... levels

**Citation Relation between documents**
- Citations related with
  - Objective
  - Methods
  - Others
Example of Structured Abstract


Abstract:

BACKGROUND: One of the major morphologic characteristics of hepatitis B is a hepatocellular regeneration which is induced by massive hepatocyte necrosis and associated with proliferative activity of hepatocytes. The purpose of this study is to document the proliferative activity of hepatocytes in various types of hepatitis B by immunohistochemical staining for proliferative cell nuclear antigen-labelling index (PCNA-LI) and electron microscopy.

METHODS: We studied 83 patients with hepatitis B: 11 cases of acute viral hepatitis, 24 cases of mild chronic hepatitis, 34 cases of severe chronic hepatitis with early cirrhosis and 14 cases of severe chronic hepatitis. The PCNA was tested by immunohistochemical staining using anti-PCNA antibody. Furthermore we evaluated the ultrastructure of acinus-forming hepatocytes (AFH) by electron microscopy. RESULTS: The expression rate and labelling index of PCNA were 27.3% and 5.3 +/- 0.9% in acute viral hepatitis, 62.5% and 22.9 +/- 31.7% in mild chronic hepatitis, and then 47.1% and 14.1 +/- 24.2% in severe chronic hepatitis with early cirrhosis, respectively (Figure 1). By contrast, no detectable PCNA expression was noted in AFH. Electron microscopic findings showed that hepatocytes forming a rosette underwent marked degenerative changes with sinusoidal capillarization and increased fine strands of collagen fiber in portal area.

CONCLUSION: The proliferative activity of hepatitis B was significantly decreased in severe chronic hepatitis containing AFH. This result suggested that differences in proliferative activity was associated with hepatic cell necrosis and AFH.
Structured Abstract Generation

From Unstructured Abstract (sentence classification)

From Main Text (sentence classification and extraction)
Preliminary experiments of sentence classification

For bio-medical domain

Context-CNN performance is the best
Lower Performance on different domain

For NLP domain

- SVM (cate): 0.37
- CNN: 0.41
- Context-CNN: 0.41

Categories:
- cate
- random
- MED
- ACL
- MIX
Vector representation of functionality of papers

1. Encoding the information of **scientific paper functionalities** (Objective, Method, Others)

   → Text

   ![Diagram of one vector representing text](Image)

   ![Diagram of three vectors](Image)

2. The correspondence of citation-related paper’s component vectors with their **citation function**
Outline of learning paper representation
Section function labeling by NLMCM rules

If the section title (in ARC) exist in NLMCM, use label from NLMCM

However ... It is impossible to label all sections ex) Parsing, MT ...

→ Predict section function label which is not labeled by NLMCM rules
The component of scientific paper and citation

Hypothesis

Citation-related papers have **at least one similar** component (Objective or Method or Other)

<table>
<thead>
<tr>
<th>3 vectors</th>
<th>most similar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective</td>
<td>0.8</td>
</tr>
<tr>
<td>Method</td>
<td>0.6</td>
</tr>
<tr>
<td>Others</td>
<td>0.2</td>
</tr>
</tbody>
</table>

**citations function** = Objective
⑤ Updating paper representation by citation graph

Update of the paper function vectors corresponding to its citation function label
# Interface for Paper Retrieval and Citation Relation

## Publication Year
- 2014 (3)
- 2013 (5)
- 2012 (1)
- 2010 (1)

## Author
- Tomas Mikolov (4)
- Christopher D. Manning (2)
- Geoffrey Zweig (2)
- Gregory S. Corrado (2)
- Jeffrey Dean (2)
- Kai Chen (2)
- Richard Socher (2)
- Wen-tau Yih (2)
- Andrew Y. Ng (1)
- Chris Dyer (1)

## Conference
- ACL (3)
- CoNLL (2)
- ArXiV (1)
- EACL (1)
- HLT-NAACL (1)
- NAACL-HLT (1)
- NIPS (1)

### OVERALL
- 10 results

### OBJECTIVE

**Linguistic Regularities in Continuous Space Word Representations**

Wen-tau Yih, Tomas Mikolov, Geoffrey Zweig

NAACL-HLT - 2013

Continuous space language models have recently demonstrated outstanding results across a variety of tasks. In this paper, we examine the vector-space word representations that are implicitly learned by the input-layer weights. We find that these representations are surprisingly good at capturing syntactic and semantic regularities in language, and that each relationship is characterized by a relation-specific vector offset. This allows vector-oriented reasoning based on the offsets between words. For example, the male/female relationship is automatically learned, and with the induced vector

### METHOD

**Efficient Estimation of Word Representations in Vector Space**

Gregory S. Corrado, Tomas Mikolov, Kai Chen, Jeffrey Dean

ArXiV - 2013

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for

**Distributed Representations of Words and Phrases and their Compositionality**

Gregory S. Corrado, Ilya Sutskever, Tomas Mikolov, Kai Chen, Jeffrey Dean

NIPS - 2013

The recently introduced continuous Skip-gram model is an efficient method for learning high-quality distributed vector representations that capture a large number of precise syntactic and semantic word relationships. In this paper we present several extensions that improve both the quality of the vectors and the training speed. By subsampling of the frequent words we obtain significant speedup and also learn more regular word representations. We also describe a simple alternative to the hierarchical softmax called negative sampling. An inherent limitation of word representations is

**Word Representations: A Simple and General Method for Semi-Supervised Learning**

Lev Art intercept, Yoav Art intercept, Joseph F. Turner

ACL - 2010

If we take an existing supervised NLP system, a simple and general way to improve accuracy is to use unsupervised word representations as extra word features. We evaluate Brown clusters, Collinbert and Weston (2009) embeddings, and HL-BL (Mnih & Hinton, 2009) embeddings of words on both NER and chunking. We use near state-of-the-art supervised baselines, and find that each of the three word representations improves the accuracy of these baselines. We find further improvements by combining different word representations. You can download our word features, for off-the-shelf use in

**Linguistic Regularities in Continuous Space Word Representations**

Wen-tau Yih, Tomas Mikolov, Geoffrey Zweig

NAACL-HLT - 2013

Continuous space language models have recently demonstrated outstanding results across a variety of tasks. In this paper, we examine the vector-space word representations that are implicitly learned by the input-layer weights. We find
Method 3: Similarity-based Knowledge Extraction

Knowledge Extraction from scientific papers

Relation extraction: Database fertilization
- Biological Science
  - KNAPSAcK Database: Plant Species-Metabolite Database
  - KEGG (Kyoto Encyclopedia of Genes and Genomes) database
- Material Science
  - Extraction of Thermoelectric elements and their properties

Data Annotation
- Annotation Tool for PDF documents
Fertilization of KNApSAcK Database

DataBases of Metabolite-Plant Species Relationship

KNApSAcK Family Databases: Integrated Metabolite–Plant Species Databases for Multifaceted Plant Research

Farit Mochamad Afendi¹,²,⁷, Taketo Okada³,⁷, Mami Yamazaki⁴,⁷, Aki Hirai-Morita¹, Yukiko Nakamura¹, Kensuke Nakamura¹, Shun Ikeda¹, Hiroki Takahashi¹, Md. Altaf-Ul-Amin¹, Latifah K. Darusman⁵, Kazuki Saito⁴,⁶ and Shigehiko Kanaya¹,⁶,*

¹Graduate School of Information Science, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma-shi, Nara, 630-0192 Japan

<table>
<thead>
<tr>
<th>Metabolite</th>
<th>Molecular Formula</th>
<th>Organism</th>
<th>Kingdom</th>
<th>Family</th>
<th>Genus</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gibberellin A1;GA1</td>
<td>C19H24O6</td>
<td>Vigna unguiculata</td>
<td>Plantae</td>
<td>Fabaceae</td>
<td>Vigna</td>
<td>Garcia-Martinez,PL (1987),212</td>
</tr>
<tr>
<td>Gibberellin A1;GA1</td>
<td>C19H24O6</td>
<td>Vitis vinifera</td>
<td>Plantae</td>
<td>Vitaceae</td>
<td>Vitis</td>
<td>Perez,Am.J.Viticul</td>
</tr>
</tbody>
</table>
Relation Annotation on Paper
Abstracts for KEGG Fertilization


Nucleotide sequence of the pyruvate decarboxylase gene from Zymomonas mobilis.
Neale AD, Scopes RK, Wettenhall RE, Hoogenraad NJ.

Abstract
Pyruvate decarboxylase (EC 4.1.1.1), the penultimate enzyme in the alcoholic fermentation pathway of Zymomonas mobilis, converts pyruvate to acetaldehyde and carbon dioxide. The complete nucleotide sequence of the structural gene encoding pyruvate decarboxylase from Zymomonas mobilis has been determined. The coding region is 1704 nucleotides long and encodes a polypeptide of 567 amino acids with a calculated subunit mass of 60,790 daltons. The amino acid sequence was confirmed by comparison with the amino acid sequence of a selection of tryptic fragments of the enzyme. The amino acid composition obtained from the nucleotide sequence is in good agreement with that obtained experimentally.

PMD: 3029726    PMCID: PMC340579
Conventional Assumption and Method for Relation Extraction

- **Supervised**
  - Entity and Relation annotated data
  - Application of supervised ML algorithms
    - Recent approaches: RNN with attention, CNN

- **Bootstrapping**
  - Starting with small seed data (entity pairs)
  - Alternated learning of patterns and entity pairs

- **Aiming at extraction of general relations**
  (commonsense)
Extraction of Scientific Relation

- Specific entities and specific relations between them
  - KNAPSAck: plant-metabolite
  - KEGG: protein-protein with enzymes/gene products
  - Databases with relations and reference papers

- Distant supervision
  - Automatic construction of training data through automatic annotation of relations in reference papers (abstracts)
Current Assumption

Existing Databases with relations and references (links to papers)

→ Distant supervision
  - Automatic construction of training data through automatic annotation of relations in reference papers (abstracts)

No database

→ Manual annotation of small scale samples
Example of entities and relation analysis

- **Span**
  - entity classification of types (e.g., species)

- **Relation**
  - entity product

Annotation

1. **Fusapyrone** and **deoxyfusapyrone**, two α-pyrone originally isolated from rice cultures of **Fusarium semitectum**, were tested in several biological assays.

2. Compounds 1 and 2 showed considerable antifungal activity against several plant pathogenic and/or mycotoxigenic filamentous fungi, although they were inactive toward yeasts isolated from plants and the Gram-positive bacterium **Bacillus megaterium** in disk diffusion assays.

3. Compound 1 was consistently more active than 2.

4. Among the tested fungi, **Fusarium** species were the least sensitive to the two pyrones, while **Alternaria alternata**, **Ascochyta rabiei**, and **species**.
Method of Relation Extraction

- **Entity detection**
  - Distant supervision + CRF, CNN-based NER

- **Extraction of novel relations**
  - Relations holding between specific entities which have never found in the DB
  - Issue: distant supervision only with (pseud-) positive examples
  - Method: Structural similarity against existing examples → kNN, One-class SVM
Similarity-based Relation Extraction

- Finding new fact via structural similarity
  - Based on representation learning at word/term/relation/expression levels
  - Possibly with integration of existing lexicon or ontology as prior knowledge
Structural similarity-based entity and relation detection

new expression

Sequence model

dependency tree model

Hopefully, joint recognition of Entities and relation
Annotation Tools

Problems of existing annotation tools

- Text data is degraded during conversion from PDF
- Some text parts may be lost, order changed, or mixed with footnotes/captions
- Some text annotation tools do not support to show section structures – original format is lost

→ We developed an annotation tool that directly make annotation of PDF file
On-going activities in our group

- **Information Structure Analysis**
  - Sentence/Section’s functionality analysis
    - Background/Objective/Method/Result/…
  - Domain-wise structure learning

- **Information Extraction from textual and non-textual parts**
  - Tables, Graphs, (Figures)

- **Recommendation system**
  - Search for similar/dissimilar articles
    - sharing the same objective, using different methods,
  - Link (reference) type analysis
Summary

- Development of Text/Document analysis tools
- Similarity-based document retrieval and recommendation
- Enabling knowledge acquisition, as well as summarization, visualization, survey
- All of those are done through various levels of structural similarity