How can Information and Communication Technology (ICT) Help us to cope with Globalization
—Text Mining as an Example—

Shigeru Masuyama
Department of Computer Science and Engineering
Toyohashi University of Technology
masuyama@tut.jp

International University Excange Programme
for Young Engineers
September 28(Tuesday) 9:00-10:30
Soukentou Seminar Room, Toyohashi University of Technology

Main Theme of this Programme:Globalization and the role of Engineers

- Era of Globalization of Economy and Mega Competition

How can engineers help us to survive under such circumstances?

This talk will exemplify a technology that is useful for such an aim.

Background(1)

- Globalization of Economy result in Global Mega Competition
- Information Explosion due to the rapid progress of Internet and Web

Screening and analyzing large amounts of information to find required one to support quick decision making are inevitable for any organzation including a company to survive and develop under such circumstances.
This is the Mission of an analyst!
A secretary or an assistant of the president play this role.

Background(2)

What analyst analyzes?

- Numerical data
  - Various statistical tools and data mining tools are widely available
- Text data
  - Manual analysis requires a lot of time and efforts

Natural language processing technology is useful for analyzing text data.
#One of the aim of research activities in our laboratory is to develop a computer that plays the role of analysts.

Background(3)

Analysts mostly analyze published(non-classified) information

The Problem is how to extract the necessary information from vast amounts of information!

- Numerical data⇒Data mining
- Text data⇒Text mining

Most of the intelligence activities are with regard to published information Analysis !

Evolution of Text Mining

- First generation(Up until the 1990s):mostly template matching
  - Example:MUC-1(87) to MUC-7(97). DARPA
- Second generation(Up to the present since 2000)
  - Detailed extraction rules created by human
  - Supervised machine learning using manually created correct data
- **Next generation**: general methods requiring few human efforts and no need to create correct data nor creating the rule

Text ming using template(First generation)

On August 18, Oyama in Miyakejima blowed up plume 8000 meters high by a large-scale eruption.

Template

Name of the event: Volcanic eruption
Name of the volcano: Oyma
Time: August 18, 2000
Place: Miyakejima Island
Kind of eruption: Plume
Scale of eruption: large-scale

Evolution of Text Mining

- First generation（Up until the 1990s）: mostly template matching
  - Example: MUC-1（87）to MUC-7（97）. DARPA
- Second generation(Up to the present since 2000)
  - Use of detailed extracted rules created by human
  - Supervised machine learning using manually created correct data
- **Next generation:** general methods requiring few human efforts and no need to create correct data nor creating the rule

A generic method that attempts to extract cause expressions for any applications

We attempt to develop a generic method that attempts to extract cause expression of events e.g., for the following applications :
1.Analysis of cause of traffic accidents
2.Analysis of cause of business performance
3.Patent mining for automatic patent map generation
4.Analysis of users views on review sites (e.g., Kakaku.com)

Text mining to support investment decisions

Hiroyuki Sakai, Shigeru Masuyama,Cause Information Extraction from Financial Articles Concerning Business Performance, IEICE Trans. Information and Systems, Vol.E91D, No.4, pp.959-968, 2008.

Introduction

- Collecting information concerning business performance is a very important task for investment.
  - If the business performance of a company is good, the stock price of the company will rise in general.
- ◆ Causal information of the business performance is also important †.
  - ■ Even if the business performance is good, the stock price will not rise if the main cause is not related to core business ††.

† In this presentation, “Causal information” denotes causal information of the business performance.
††This is also the case for the bad business performance.

Purpose

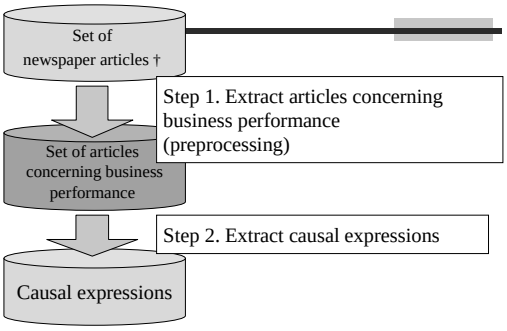
We propose a method of extracting causal information from Japanese newspaper articles concerning business performance.

Causal expression
a phrase implying causal information

Examples of Causal expressions

- *zidousya no uriage ga koutyou* (Car sales were good)
- *kokunai no zigyou no saikoutiku ga soukou* (Domestic restructuring was successful)

Overview of our method



†Nikkei newspapers published from 2001 to 2005

Extraction of articles concerning business performance

- Our method extracts articles concerning business performance by using Support Vector Machine.
  - Training data
    - [Positive examples :] 2,920 articles concerning business performance.
    - [Negative examples :] 2,920 articles not concerning business performance.
  - Features

    Content words contained in the positive examples

20,880 articles concerning business performance are extracted†.

† from Nikkei newspapers published from 2001 to 2005

Causal expression

Causal expression

a phrase implying causal information contained in a sentence consisting of some "*bunsetsu*".

a "*bunsetsu*" is a basic block in Japanese composed of several words

Our method extracts causal expressions by using clue phrases, which are phrases frequently modified by causal expressions.

Examples of causal expressions	Clue phrase
• *zidousya no uriage ga koutyou* (Car sales were good)	*ga koutyou* (are good)
• *hon no uriage ga husin* (Computer sales were down)	*ga husin* (are down)

Clue phrase

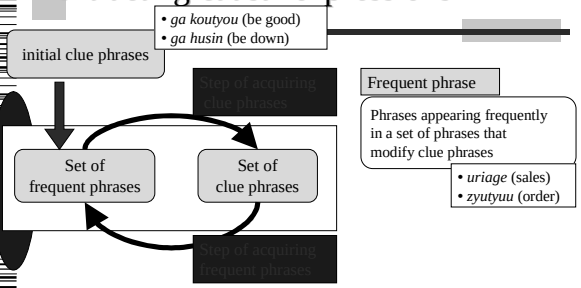
Examples of clue phrases

de oginau (cover), *ga zyuntyou* (go well)
ga kiyo (contribute), *ga kentyou* (is robust)

- If many clue phrases effective for extracting causal expressions is possible to acquire, causal expressions are extracted automatically.
- It is hard to acquire sufficient clue phrases effective for extracting causal expressions by hand.

Our method also acquires such clue phrases automatically from a set of articles concerning business performance.

Overview of our method for extracting causal expressions



Frequent phrase

Phrases appearing frequently in a set of phrases that modify clue phrases

- *uriage* (sales)
- *zyutyuu* (order)

Step 1. Input a few initial clue phrases and acquire frequent phrases.

Frequent phrase

Definition

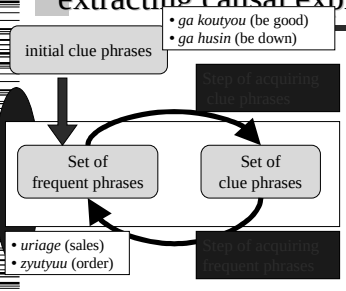
Phrases appearing frequently in a set of the phrases that modify clue phrases

Examples of frequent phrases

- *zidousya no uriage ga koutyou* (car sales were good)
- *hon no uriage ga husin* (book sales were down)
- *konpuuter no uriage ga otikonda* (computer sales were weak)

Frequent phrase : *uriage* (sales)
Blue characters : clue phrases

Overview of our method for extracting causal expressions



Step 2. Acquires new clue phrases by using frequent phrases.

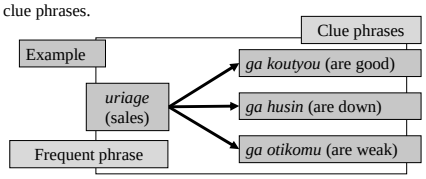
New clue phrases
• *zouka* (be increasing)

Step 3. Repeats "Step of acquiring clue phrases" and "Step of acquiring frequent phrases" until they are executed predetermined times.

Screening of frequent phrases

Our method selects appropriate frequent phrases.

- An appropriate frequent phrase is one that modifies various kinds of clue phrases.



- Our method calculates entropy based on the probability that a frequent phrase modifies a clue phrase.
- If a frequent phrase modifies various kinds of clue phrases, the entropy is large.

Entropy $H(e)$

$$H(e) = -\sum_{s \in S(e)} P(e,s) \log_2 P(e,s)$$

$P(e,s)$: the probability that frequent phrase e modifies clue phrase s.
$S(e)$: the set of clue phrases modified by frequent phrase e.

Threshold value

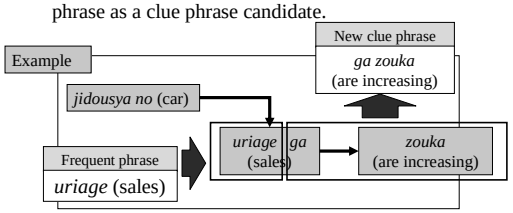
$$T_e = \alpha \log_2 |Ns|$$

α : constant ($0 < \alpha < 1$)
Ns : the set of clue phrases used for extracting frequent phrases

Our method selects frequent phrases assigned entropy $H(e)$ larger than threshold value T_e.

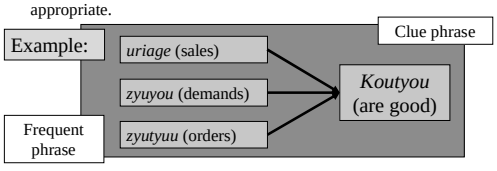
Acquisition of new clue phrases

- Our method acquires new clue phrases by using frequent phrases.
  - Our method extracts a phrase modified by a frequent phrase as a clue phrase candidate.



Screening of clue phrases

- Our method selects appropriate clue phrases.
  - A clue phrase modified by various kinds of frequent phrases is appropriate.



Step 1. Calculate entropy based on the probability that a clue phrase is modified by a frequent phrase.

Step 2. Select clue phrases assigned the entropy larger than a threshold value as appropriate clue phrases.

Elimination of inappropriate clue phrases (1)

Our method eliminates inappropriate clue phrases.

Example		
	appropriate	*ga koutyou* (are good)
	inappropriate	*Ni naru* (become)

Definition:

[*Sp*:] the set of articles concerning business performance
[*Sn*:] the set of articles not concerning business performance

Our method eliminates inappropriate clue phrases by using statistical information of *Sp* and *Sn*

Elimination of inappropriate clue phrases (2)

- We use the following rules for eliminating inappropriate clue phrases

Observations:

i. Appropriate clue phrases frequently appear in *Sp*

ii. Inappropriate clue phrases frequently appear in not only *Sp* but also *Sn*.

- A clue phrases that frequently appear in only *Sp* is not eliminated.
- A clue phrase that frequently appears in not only *Sp* but also *Sn* is eliminated.

[*Sp*:] the set of articles concerning business performance

[*Sn*:] the set of articles not concerning business performance

Conclusion

- Our method is able to extract causal expressions appropriately.
- Our method extracts causal expressions without predetermined patterns or complex rules given by hand.
- It is expected to be applied to other tasks or other language for acquiring phrases that have a particular meaning not limited to causal information.

Related work

- Adding polarity to a causal expression.
  - "Car sales are good" : [Positive]
  - "Computer sales are down" : [Negative]

(Hiroyuki Sakai, Shigeru Masuyama, Assigning polarity to causal Information in Finantial Articles on Business Performance of Companies, IEICE Trans. Information and Systems, E92-D, 12, pp.2341-2350, 2009.)

- Identifying whether a causal expression is related to core business of the company or not.
  - "Car sales are good" : [Related]
  - "Profit from sales of stocks" : [Not related]

Polarity（positive, negative）assignment to cause information on business performance

1. Assign polarity to each document (an article on economy) in the document set (set of articles in economy).
2. We assign polarity to a causal expression with regard to which document set, positive or negative, the expression is included.

Classification of a document is done by

Identifying whether a causal expression is related to core business of the company or not

- Are the performance of a company related to that of core business ?（Fujimura et al 09）

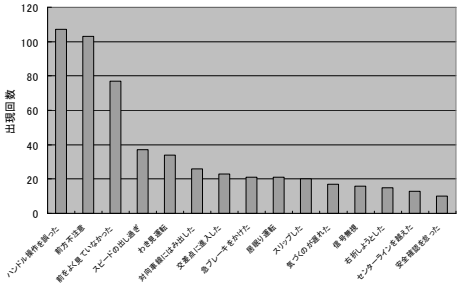
An example of cause expression related to the core business

Sector of external sales of food and meat was stronger than expected

An example of cause expression not related to the core business

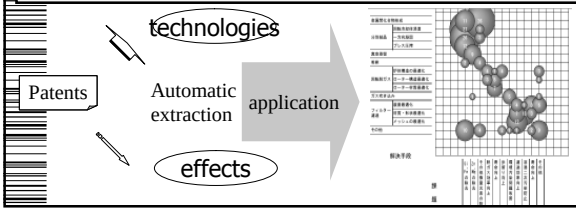
Extraordinary income of 83 billion yen was recorded from the gain on the sale of abandoned housing

Extraction of cause expressions for traffic accidents



Expression extraction from patents and its application to automatic patent map generation

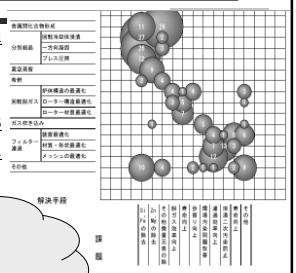
Extracting technology expressions and effect expressions are musts for automatic generation of patent maps



Background

Patent map visualizes the trend of patent applications

1. Easy to understand as a classification of technologies
2.Patent examiners consider Technologies and their effects when they examin applied patents



The proposed method is useful to automatic generation of patent maps !

Patent map

- Required at the corporate intellectual property department and the research and development section in a company to analyze and devise research and development and intellectual property strategy
- Decide whether to develop new technology or to purchase the rights by examining the status of rights needed to develop the new product
  - Currently, they have been created by hand and takes much labor and cost.
- ⇒ Aims to establish an automatic patent map generation method applying automatic extraction technology and semantic representation of a causal relationship that have been developed in our laboratory

Tecnologies vs effect expressions

- Effects（effect expressions）
  - Direct user benefits
- Technologies（technology expressions）
  - technologies to realize the direct benefits

Example

Technology expressions

According to the present invention, prevention of the adhesion of the adhesive substance may settle to a minimum maintenance.

Effect expressions

clue expression

Thank you for your kind attention!