5-5 Fundamental Language Resources

HASHIMOTO Chikara, OH Jong-Hoon, SANO Motoki, and KAWADA Takuya

Fundamental language resources are classified into natural language processing tools and natural language data, which are used as building blocks for natural language information processing systems such as question answering systems and information analysis systems. Various kinds of natural language information processing systems generally have necessary fundamental language resources in common. However, some fundamental language resources are difficult to construct for some organizations due to limited computational capability, limited manpower, budget constraint, or time constraint. Thus, it is important to construct and publish such fundamental language resources in order for the research community to make steady progress. We, Information Analysis Laboratory members, have constructed and published many fundamental language resources that are precise and have wide-coverage, some of which are difficult to construct for some organizations, with a large-scale high-performance computing environment, many researchers who are acquainted with natural language processing, and many richly-experienced linguistic data annotators. In this paper, we present fundamental language resources that we have constructed, including those that will be released in the near future. We do not present natural language processing tools that have described in **5-4** of this special issue.

Keywords
Language resources, Dictionaries, Corpora, Language processing tools, ALAGIN Forum

1 Introduction

In the midst of the information explosion era, natural language information processing systems such as question answering systems and information analysis systems that can perform precise retrieval of required information from the so called Big Data have definitely increased in importance. Such language processing systems often require a high level of "language comprehension" ability. For example, when a question answering system receives a question such as "What can be caught in the Kawazugawa River?", the system is supposed to find the answer by detecting the candidate sentences that do not contain the phrase "can be caught in the Kawazugawa River" such as "The sweetfish fishing season has come to the Kawazugawa River", "Marbled eels live in the Kawazugawa River" or "Beautiful dwarf rill trout in the Kawazugawa River" to retrieve "sweetfish", "marbled eels" and "dwarf rill trout" as the right answers. As human language comprehension is believed to be supported by the ability to analyze sentences by using various linguistic knowledge, computers also need a wide range of linguistic knowledge (language data) and an analyzer (language processing tool) in order to understand language. In this article, we generically call such language data and processing tools "fundamental language resources".

In general, fundamental language resources used as necessary building blocks for the construction of high-performance natural language information processing systems include a wide range of systems. Moreover, construction of a language resource requires not only

technology, experience and expertise but a huge cost for securing necessary resources such as large-scale computing environment and manpower. Therefore, some organizations cannot afford to construct a necessary fundamental language resource without external support, which hinders the research community from making steady progress as a whole.

Universal Communication Research Institute's Information Analysis Laboratory constructs and provides highly precise language resources by utilizing its rich assets including a huge collection of Web-extracted text data, a large-scale parallel computing environment, many richly-experienced linguistic data annotators and researchers with expertise in information processing. With the aim of making steady progress as a whole research community, we have constructed and published a large number of fundamental language resources. Some of them are essential for various language information processing systems such as question answering systems and information analysis systems, which are very expensive to construct.

In this paper, we present fundamental language resources that we have constructed, including those that have not been published yet. We do not present natural language processing tools described in **5-4** in this special issue [1].

Tables 1 and 2 provide the list of fundamental language resources that are presented in Section **2** and succeeding sections. Table 1 contains the fundamental language resources that are available only to the members of Advanced LAnGuage INformation Forum (ALAGIN[*1]), and those in Table 2 are freewares available to the public. The terms "DB", "Service" and "Tool" under the column "Type" represent "database", "Web-based service" and "tool" respectively.

2 Advanced LAnGuage INformation Forum: ALAGIN

Advanced LAnGuage INformation Forum (ALAGIN) is a forum that aims to disseminate and promote the technologies for realizing a highly advanced form of communication where language differences pose no barrier. Since its establishment in 2009, the forum has been bringing together knowledge and expertise of researchers from industries, academia, research institutions and the government for conducting its researches including the development of text and speech translation systems, spoken dialogue systems, and information analysis and advanced information retrieval technologies for retrieving desired information or judging the credibility of the acquired information. Moreover, the researchers have been developing, testing and standardizing an unprecedented size of language resources (e.g. dictionaries and corpora) that are necessary for developing the above mentioned technologies, aiming to provide the resulting tools and language resources for the forum members.

The language resources presented in this paper and the natural language processing tools presented in **5-4** in this special issue [1] including freewares are available on ALAGIN's language resource distribution site[*2].

ALAGIN also provides the tools and data that have been developed and constructed by Universal Communication Research Institute's Multilingual Translation Laboratory and Spoken Language Communication Laboratory.

For further details of ALAGIN including its activities and size, please see **8-1** [2] in this special issue.

3 Databases of semantic relations between nominals

3.1 Case Base for Basic Semantic Relations

"Case Base for Basic Semantic Relations" contains 102,436 pairs of nominals manually classified and annotated with semantic relations. The entry pairs had been chosen from among approximately one hundred million pages of Web documents based on the similarity of the contexts where the potential entry

＊1 and ＊2 http://alaginrc.nict.go.jp/

Table 1 *List of language resources: Available only to ALAGIN members*

Name	Published Year	Type	Size
Database of Japanese Paraphrasing Patterns	2009	DB	approx. 2.5 billion entries
Verb Entailment Database	2009	DB	approx. 120K pairs
List of Burden and Trouble Expressions	2009	DB	approx. 20K entries
Database of Similar Context Terms	2009	DB	approx. 1 million entries
Hypernym Hierarchy Database	2009	DB	approx. 700K entries
Database of Word Co-occurrence Frequency	2009	DB	approx. 1 million entries
Support Service for Customized Word Set Generation	2010	Service	—
Japanese Dependency Structure Database	2010	DB	approx. 4.6 billion entries
Case Base for Basic Semantic Relations	2010	DB	approx. 100K entries
Database of Japanese Orthographic Variant Pairs	2010	DB	approx. 1.6 million entries
Semantic Relation Acquisition Service	2011	Service	—
Kyoto Sightseeing Blogs for Evaluative Information	2011	DB	approx. 1K articles
Predicate Phrase Entailment Database	will be published around the end of FY 2012	DB	approx. 600K pairs
Excitatory/Inbitory Template Databas	will be published around the end of FY 2012	DB	approx. 10K entries
Predicate Phrase Contradiction Database	will be published around the end of FY 2012	DB	approx. a million pairs
Predicate Phrase Causality Database	will be published around the end of FY 2012	DB	approx. a million pairs

Table 2 *List of language resources: Freewares*

Name	Published Year	Type	Copyright & License	Size
Japanese WordNet	2009	DB	constructed by NICT	approx. 90 K words
Hyponymy Extraction Tool	2010	Tool	GPL	—
Dependency Structure Database of Japanese Wikipedia Entries	2011	DB	CC BY-SA 3.0	approx. 8 hundred million entries
Para-SimString	will be published by the end of FY 2012	Tool	Modified BSD, LGPL, or GPL	—
QE4Solr	will be published by the end of FY 2012	Tool	Modified BSD, LGPL, or GPL	—

words were contained [3]. For example, a pair "電子計算機 (*denshikeisanki*) / computer" and "電算機 (*densanki*) / computer" is classified as a abbreviation pair, and a pair "患部 (*kanbu*) / affected part" and "治療部位 (*chiryobui*) / affected part" is classified as a synonym pair. Table 3 lists all semantic relations used in the database [4].

Notational variant pairs have the same pronunciation and meaning but different transcription patterns such as "問い合わせ (*toiawase*) / inquiry — 問合せ (*toiawase*) / inquiry", abbreviation pairs have the same meaning but one is the abbreviation or shortened form of the other such as "つくばエクスプレス / Tsukuba Express — TX", synonym pairs are those that denote the same thing or phenomenon but cannot be classified neither as orthographic variant nor abbreviation pairs, such as "乳飲み子 / infant — 赤ン坊 / baby", contrastive term pairs are those contrast with each other, such as "乾麺 / dried noodles — 生麺 / fresh noodles", in meronym pairs, one term is a part of the other, either physically or conceptually, such as "たし算 / addition — 四則計算 / four arithmetic operations", and collocational pairs have the same super-ordinate which is not too abstract, such as "にわか雨 / sudden rain shower — 夕立 / late afternoon shower".

Table 3 *Semantic relation categories in "Case Base for Basic Semantic Relations"*

Category	Example pairs
Notational variant	問い合わせ (*toiawase*) "inquiry" — 問合せ (*toiawase*) "inquiry"
Abbreviation	つくばエクスプレス (*tsukubaekusupuresu*) "Tsukuba Express" — TX "TX"
Synonym	乳飲み子 (*chinomigo*) "infant" — 赤ン坊 (*akanbo*) "baby"
Contrastive	乾麺 (*kanmen*) "dried noodles" — 生麺 (*namamen*) "fresh noodles"
Meronym	たし算 (*tashizan*) "addition" — 四則計算 (*shisokukeisan*) "four arithmetic operations"
Collocational	にわか雨 (*niwakaame*) "sudden rain shower" — 夕立 (*yudachi*) "late afternoon shower"

What makes "Case Base for Basic Semantic Relations" unique is its wide coverage. It contains a number of pairs related by certain relationships, and those terms include not only common nouns but proper nouns and technical terms that are hardly listed in commonly used thesauruses. For example, its synonym pairs include "サイテス / CITES and ワシントン条約 / Washington Convention", "サンフランシスコ講和条約 / San Francisco Peace Treaty and 対日講和条約 / Treaty of Peace with Japan", "シナイ山 / Mount Sinai and ホレブ / Horeb", "バックカントリースキー / backcountry skiing and 山スキー / off-piste skiing", and "シナジー効果 / synergy effect and 相乗効果 / synergy effect". This database can be utilized for retrieving larger numbers of information by adding, for instance, "サイテス / CITES" as an additional search word to the word "ワシントン条約 / Washington Convention" input by a user.

3.2 Database of Japanese Orthographic Variant Pairs

Database of Japanese Orthographic Variant Pairs contains positive and negative instances of Japanese orthographic variant pairs (or pairs of orthographically inconsistent terms). Examples of orthographic variant pairs for the term "ギョウザ (*gyoza*) / gyoza dumpling" include "ギョウザ — ギョーザ", "ギョウザ — ぎょうざ", and "ギョウザ — 餃子" ('—' is used for indicating a boundary between two terms in these examples). One of the typical application usages of an orthographic variant database is query expansion in information retrieval operations. For example, when a user inputs the search word "餃子 (*gyoza*) / gyoza dumpling", the search system can automatically expand the search criteria to "餃子 OR ギョーザ OR ギョウザ OR ぎょうざ".

The database contains only the term pairs in which only one character is different (i.e. the edit distance between the two terms is one). Orthographic variant pairs whose edit

distance value is larger than 1, like “ギョーザ — 餃子 ”, are not listed in this database.

Note that “Case Base for Basic Semantic Relations” presented in Subsection **3.1** does not apply this edit distance-based constraint to its orthographic variant pairs (the number of pairs listed as orthographic variant pairs is about 30,000), while Database of Japanese Orthographic Variant Pairs contains more than a million pairs satisfying the constraint..

The following are examples of orthographic variant pairs listed in Database of Japanese Orthographic Variant Pairs.

- “Center — center” (higher and lower cases)
- “ゴミ置き場 — ゴミ置場 (*gomiokiba — gomiokiba*) / a garbage collection point” (different usages of declensional kana endings)
- “ギタープレー — ギタープレイ (*gitapure — gitapurei*) / guitar playing” (difference of “ー” and “イ” at the end of each word)
- “ツィンマーマン — ツィマーマン (*tsuinmaman — tsuimaman*) / Zimmermann” (lack of “ン” in the latter)
- “ブルース・スプリングスティーン — ブルーススプリングスティーン (*burusu supuringusutein — burususupuringusutein*) / Bruce Springsteen” (lack of “・” in the latter)

For constructing Database of Japanese Orthographic Variant Pairs, we first prepared orthographic variant pairs by using the method proposed by Kuroda et al. [4]. This manually prepared data consist of 48,067 pairs of orthographic variants, 10,730 pairs of semi-orthographic variants and 2,758 synonyms (not orthographic variants). Table 4 shows the

Table 4 *Examples of manually prepared orthographic variant pairs*

Types	Examples
Orthographic variant pairs	“第一週目 ― 第 1 週目 (*daiisshume ― daiisshume*) / first week”, “4 カ月後 ― 四カ月後 (*yonkagetsugo ― yonkagetsugo*) / 4 months later”, “Flash Player ― Flash player”, “Center ― center”, “ゴミ置き場 ― ゴミ置場 (*gomi okiba ― gomi okiba*) / garbage collection point”, “割引き価格 ― 割り引き価格 (*waribiki kakaku ― waribiki kakaku*) / discount price”, “ギタープレー ― ギタープレイ (*gita pure ― gita purei*) / guitar playing”, “ブルース・スプリングスティーン ― ブルーススプリングスティーン (*burusu supuringusutein ― burusu supuringusutein*) / Bruce Springsteen”
Semi orthographic variant pairs	“法違反 ― 法律違反 (*ho ihan ― horitsu ihan*) / violation of law”, “補足給付 ― 補足的給付 (*hosoku kyufu ― hosokuteki kyufu*) / supplementary benefit”, “調査法 ― 調査手法 (*chosa ho ― chosa shuho*) / investigation method”, “株取得 ― 株式取得 (*kabu shutoku ― kabushiki shutoku*) / stock acquisition”, “米本社 ― 米国本社 (*bei honsha ― beikoku honsha*) / US headquarters”, “手数料額 ― 手数料金額 (*tesuryo gaku ― tesuryo kingaku*) / amount of fee”, “胴体下 ― 胴体下部 (*dotai shita ― dotai kabu*) / belly compartment”, “満州軍 ― 満州国軍 (*manshu gun ― manshukoku gun*) / Manchukuo Imperial Army”, “土曜・日曜 ― 土曜・日曜日 (*doyo nichiyo ― doyo nichiyobi*) / Saturday and Sunday”, “依頼者 ― 依頼者様 (*iraisha ― iraisha sama*) / client”
Synonyms (non orthographic variant pairs)	“コンスタンティヌス ― コンスタンティヌス帝 (*konsutanteinusu ― konsutanteinusu tei*) / Roman Emperor Constantine”, “インテル ― インテル社 (*interu ― interu sha*) / Intel”, “シックスアパート ― シックスアパート社 (*shikkusuapato ― shikkusuapato sha*) / Six Apart Ltd.”, “米アップル ― 米アップル社 (*bei appuru ― bei appuru sha*) / Apple Inc. US”, “Siemens ― Siemens 社 (*shimensu ― shimensu sha*) / Siemens AG”, “フィナンシャル・タイムズ ― フィナンシャル・タイムズ紙 (*finansharutaimuzu ― finansharutaimuzu shi*) / the Financial Times”, “ビハール ― ビハール州 (*biharu ― biharu shu*) / State of Bihār”, “北海道札幌 ― 北海道札幌市 (*hokkaido sapporo ― hokkaido sapporo shi*) / Sapporo, Hokkaido”, “差別的 ― 差別的だ (*sabetsuteki ― sabetsuteki da*) / being discriminative”, “エリア外 ― エリア以外 (*eria gai ― eria igai*) / outside the service area”

examples. Then we automatically acquired orthographic variant pairs from 100 million Web documents by using the method proposed by Kojima et al. [5]. We first extracted the 10 million most frequent words and phrases (mostly, words) from 100 million Web documents and selected as the final candidates only the orthographic variant pairs with "edit distance 1" from all the possible combinations of the 10 million words and phrases. Then, we classified these candidates into orthographic variant pairs and non-orthographic variant pairs by using a classifier trained by using our manually prepared data described above. Finally, 1.15 million to 1.53 million pairs of orthographic variants were automatically acquired with the precision rate of 95% or higher.

3.3 Database of Similar Context Terms

Database of Similar Context Terms contains about 1 million sets of nouns and noun phrases. Each set has one entry word and its most similar 500 terms (nouns or noun phrases) in contexts extracted from 100 million

Table 5 *Examples of similar context terms*

Entry	Similar context terms
ルパン三世 (rupansansei) / Lupin the Third (a Japanese animation)	ルパン 3 世 (rupansansei) / Lupin the Third (-0.229), 名探偵コナン (*meitantei konan*) / Detective Conan (-0.259), 宇宙戦艦ヤマト (*uchusenkan yamato*) / Space Cruiser Yamato (-0.265), ケロロ軍曹 (*keroro gunso*) / Sergeant Frog (-0.28), 鉄腕アトム (*tetsuwan atomu*) / Mighty Atom (-0.282), ガッチャマン (gacchaman) / Gatchaman (-0.287), デビルマン (debiruman) / Devil Man (-0.289), サイボーグ 009 (*saibogu 009*) / Cyborg 009 (-0.294), 新世紀エヴァンゲリオン (*shinseiki evangerion*) / Neon Genesis Evangelion (-0.295), ヤッターマン (yattaman) / Yatterman (-0.305), 聖闘士星矢 (*seinto seiya*) / Saint Seiya (-0.308), セーラームーン (*seramun*) / Sailor Moon (-0.308) ...
チャイコフスキー (*chaikofusuki*) / Tchaikovsky	ブラームス (*buramusu*) / Brahms (-0.152), シューマン (*shuman*) / Schumann (-0.163), メンデルスゾーン (*menderusuzon*) / Mendelssohn (-0.166), ショスタコーヴィチ (*shosutakovichi*) / Shostakovich (-0.178), シベリウス (*shiberiusu*) / Sibelius (-0.18), ハイドン (*haidon*) / Haydn (-0.181), ヘンデル (*henderu*) / Handel (-0.181), ラヴェル (*raveru*) / Ravel (-0.182), シューベルト (*shuberuto*) / Schubert (-0.187), ベートーヴェン (*betoven*) / Beethoven (-0.19), ドヴォルザーク (*dovoruzaku*) / Dvorak (-0.192), ラフマニノフ (*rafumaninofu*) / Rachmaninov (-0.193), バルトーク (*barutoku*) / Bartok (-0.198) ...
カラヤン (*karayan*) / Karajan	クレンペラー (*kurenpera*) / Klemperer (-0.21), バーンスタイン (*bansutain*) / Bernstein (-0.215), トスカニーニ (*tosukanini*) / Toscanini (-0.227), フルトヴェングラー (*furutovengura*) / Furtwangler (-0.227), ベーム (*bemu*) / Boehm or Böhm (-0.23), チェリビダッケ (*cheribidakke*) / Celibidache (-0.232), アバド (*abado*) / Abbado (-0.239), ムラヴィンスキー (*muravinsuki*) / Mravinsky (-0.242), クーベリック (*kuberikku*) / Kubelik (-0.245), ヴァント (*vanto*) / Wand (-0.254), リヒテル (*rihiteru*) / Richter (-0.256), メンゲルベルク (*mengeruberuku*) / Mengelberg (-0.256), ハイティンク (*haitinku*) / Haitink (-0.265), アーノンクール (*anonkuru*) / Harnoncourt (-0.276) ...
ストーンズ (*sutonzu*) / The (Rolling) Stones	YMO (-0.215), メタリカ (*metarika*) / Metallica (-0.223), ビートルズ (*bitoruzu*) / The Beatles (-0.236), ローリング・ストーンズ (*roringu sutonzu*) / The Rolling Stones (-0.245), エアロスミス (*earosumisu*) / Aerosmith (-0.268), ツェッペリン (*tsuepperin*) / (Led) Zeppelin (-0.277), Beatles (-0.284), ローリングストーンズ (*roringusutonzu*) / The Rolling Stones (-0.287), クイーン (*kuin*) / QUEEN (-0.292), ベンチャーズ (*benchazu*) / The Ventures (-0.294), ビーチ・ボーイズ (*bichi boizu*) / The Beach Boys (-0.295), ピンク・フロイド (*pinku furoido*) / Pink Floyd (-0.297), レッド・ツェッペリン (*reddo tsuepperin*) / Led Zeppelin (-0.301), ラモーンズ (*ramonzu*) / Ramones (-0.301), ディープ・パープル (*dipu papuru*) / Deep Purple (-0.301), ニール・ヤング (*niru yangu*) / Neil Young (-0.305), ザ・フー (*za fu*) / The Who (-0.306) ...

Web documents. Table 5 shows the examples. In these examples, the scores following each term represent its contextual similarity to the given term. You can see that the title of animation movies and TV shows are chosen as terms having similar contexts with a famous Japanese animation "Lupin the Third", famous composers are listed for "Tchaikovsky", celebrated conductors for "Karajan" and old-time rock bands for "The (Rolling) Stones".

These similar context terms have been proved to be effective in several natural language processing tasks including acquisition of semantic relations such as causal relationship [6] and question answering tasks for the "why" questions [7]. For example, the preferred answers for a question that asks the cause of a disease like "What causes cancer?" often include the names of toxic substances, viruses and body parts that are related to the disease in the question. In other words, when a question includes the word "cancer" or its similar words (i.e. similar context terms of "cancer"), candidate sentences in the correct answers tend to contain the similar context terms of words that represent "a toxic substance", "a virus" and "a body part". This database enables us to capture such tendency in the relationship between a question and the correct answers and thus allows us to improve the performance of question answering systems.

For the details of automatic acquisition of similar context terms, please see the references [3], [8] and [9] written by Kazama et al. The contexts of the documents used for the construction of the database are also presented in Subsection **5.1** of this paper.

3.4 Hypernym Hierarchy Database

Hypernym Hierarchy Database is a hierarchical thesaurus containing approximately 69,000 nouns and noun phrases. We have manually built a set of hierarchies between the hypernyms in hyponymy relation (hypernym/hyponym pairs), that are automatically acquired from Japanese Wikipedia articles (ver. 2007/03/28) by using the "Hyponymy Extraction Tool" presented in Subsection **6.1**. The hierarchy between hypernyms enables us to estimate semantic association between automatically acquired hypernym/hyponym pairs. For example, the hypernyms in the hypernym/hyponym pairs "黒澤明の映画作品 (movie work by Akira Kurosawa) → 七人の侍 (Seven Samurai)" and "映画作品 (movie work) → ローマの休日 (Roman Holiday)" can be hierarchized as below:

- 作品 (work) → 映画作品 (movie work) → 黒澤明の映画作品 (movie work by Akira Kurosawa)
- 作品 (work) → 映画作品 (movie work)

This means that "七人の侍 (Seven Samurai)" and "ローマの休日 (Roman Holiday)" have the same hypernym "映画作品 (movie work)", which helps us to estimate that these two terms may belong to the same concept class (i.e. movie).

To build a hierarchy between hypernyms, we first morphologically analyzed hypernyms in hyponymy relations acquired by using the Hyponymy Extraction Tool and extracted head nouns or head noun phrases of the hypernyms. For example, "黒澤明の映画作品 / movie work by Akira Kurosawa" in the above example has three Japanese head noun or noun phrases, "作品 (work)", "映画作品 (movie work)" and "黒澤明の映画作品 (movie work by Akira Kurosawa)." These head nouns or noun phrases are then manually checked whether they can serve as a hypernym of the hypernym in a given hyponymy relations. For the details of building a hierarchy between hypernyms, please see the paper Kuroda et al. [10]. This database has been proven to be effective in a task of linking hyponymy relations extracted from Wikipedia articles to Japanese WordNet [11][*3].

*3 According to the Reference [12] by Kuroda et al., the matching ratio between the hypernyms acquired from Wikipedia articles and the WordNet synset had been as low as 8% at the beginning, but after the introduction of this database, the ratio became as high as 95%.

3.5 Database of Word Co-occurrence Frequency

Database of Word Co-occurrence Frequency consists of a collection of co-occurring word lists. Each list has an entry word and co-occurring words that are semantically related to the entry. Their semantic relationship was estimated by three different measures, Dice coefficient, DPMI [13] and co-occurrence frequency. These three measures were calculated by using co-occurrence frequencies in a 100 million Web documents with the following three different conditions:

- ○ Co-occurrence in a document between all combinations of approx. 1 million words.
- ○ Co-occurrence within 4 neighboring sentences between all combinations of approx. 0.5 million words.
- ○ Co-occurrence in a sentence between all the combinations of approx. 0.5 million words.

Since words with a strong semantic association with each other tend to co-occur, Word Co-occurrence Frequency Database can be used as an associated word database. For example, the top-5 words of "Christmas" and "baseball" by Dice co-efficient in this database are as follows:

クリスマス (*kurisumasu*) / Christmas: "お正月 (*oshogatsu*) / New Year Day" (0.172339), "誕生日 (*tanjobi*) / birthday" (0.119606), "サンタ (*santa*) / Santa Claus" (0.113987), "冬(*fuyu*) / winter" (0.112612), "年末 (*nenmatsu*) / year end" (0.110775)

野球 (*yakyu*) / Baseball: "サッカー (*sakka*) / soccer" (0.362974), "格闘技 (*kakutogi*) / combat sport" (0.227781), "プロ野球 (*puro-yakyu*) / professional baseball" (0.220464), "ゴルフ (*gorufu*) / golf" (0.210349), "テニス (*tenisu*) / tennis" (0.208742)

Word Co-occurrence Frequency Database has been proved to be effective in its usage for the analogy-based acquisition of semantic relations between words [14].

3.6 List of Burden and Trouble Expressions

"List of Burden and Trouble Expressions" is a database containing 20,115 expressions related to troubles and obstacles that may be a burden on human activities or have a negative impact, such as "disaster", "psychological stress" and "asbestos contamination". The trouble and burden related expressions in the database were automatically acquired from Web documents based on the method proposed by De Saeger et al. [15] and manually checked and classified. The expressions are annotated with category labels such as "disease", "suffering", "illegal act / violation" and "hazardous substance". For example, "hepatitis B", "influenza" and "cryptococcosis" are classified as "disease", "chemical accident", "herbivory in coral reefs" and "thalidomide incident" as "suffering", "skimming", "falling asleep while driving" and "infringement of rights" as "illegal act / violation" and "sleeping gas", "acid precipitates" and "vehicle emission" as "hazardous substance" respectively. Table 6 shows other examples of trouble and burden expression labels and their examples.

Construction of a large scale list of burden and trouble expressions enables a comprehensive search of unexpected troubles. One exam-

Table 6 *Examples of burden and trouble expressions*

Category	Examples
Error	core dump / core dump, DB エラー (*DB era*) / DB error, Out of Memory / Out of Memory, アンダーフロー (*andafuro*) / underflow
Natural phenomenon	エルニーニョ (*eruninyo*) / El Nino, かまいたち (*kamaitachi*) / whirlwind, メイルシュトローム (*meirushutoro-mu*) / maelstrom, 黄砂 (*kosa*) / yellow dust
Physical damage	メルトダウン (*merutodaun*) / meltdown, ラインブレイク (*rainbureiku*) / line break, 液晶割れ (*ekishoware*) / LCD cracking, 荷痛み (*niitami*) / damage during handling and transporting
Harmful organism	レタス病害虫 (*retasubyogaichu*) / lettuce pests and diseases, アオコ (*aoko*) / algae bloom, アクネ菌 (*akunekin*) / propionibacterium acnes, ネキリムシ (*nekirimushi*) / cutworm

ple is a search of burden and trouble expressions related to the Great East Japan Earthquake on the social networking site, Twitter. We searched about 3.2 million tweets related to the earthquake posted during the period from March 11th to June 17th of 2011 [16] for burden and trouble expressions, and identified not just comments related to common predictable troubles such as "power failure" and "water supply suspension" but those related to "disaster related death" or troubles resulting from a secondary disaster such as "carbon monoxide poisoning" caused by briquettes used for surviving the cold weather during the lifeline suspension and "economy class syndrome" due to living and sleeping in a car instead of staying in a public safe shelter. In this way, the list of more than 20,000 burden and trouble expressions is usable in identifying unpredictable troubles.

3.7 Japanese WordNet

Inspired by Princeton University's Princeton WordNet and other like resources, the Japanese WordNet was developed to classify Japanese words into groups called "synsets". A synset is a group of words that have the same concept, and currently, 93,834 words are contained in the Japanese WordNet. For example, words like "行動 / behavior", "営み / work", "行為 / behavior", "活動 / activity" and "営為 / deed" are put into a group (synset ID: 00030358-n) with its definition "for human beings to do something or to start doing something" and a usage example "殺人と他の異常な行動の話があった / We heard a story about murder and other abnormal behaviors". The Japanese WordNet also has some verbs and adjectives besides nominals.

Besides grouping words into synsets of synonyms, the Japanese WordNet provides information on semantic relations such as hypernym relations (e.g. "furniture — chair") and meronym relations (e.g. "leg — chair"). Some semantic relation links used in the Japanese WordNet and their examples are shown Table 7.

The link "Hypernym" relates a pair of synsets where the concept represented by one synset is the hypernym of that of the other such as "animal — poikilotherm". The link "Meronyms" is for a pair of synsets where one is a constituent of the other such as "automobile — airbag". The link "Causes" relates a pair of synsets where the occurrence or existence of one synset prompts that of the other such as "project (a film) — appear". The link "Entails" relates a pair where the existence of the event represented by one synset means the simultaneous or preceding occurrence of the event represented by the other such as "overcharge — request". The links "Causes" and "Entails" are further explained in Subsections **4.5** and **4.1** respectively.

Table 7 *Relation links used in Japanese WordNet and their examples*

Link	Example
Hypernym	動物 (*dobutsu*) / animal — 変温動物 (*henondobutsu*) / poikilotherm
Meronyms	エアバック (*eabakku*) / airbag — 自動車 (*jidosha*) / automobile
Causes	映写する (*eishasuru*) / project — 表れる (*arawareru*) / appear
Entails	吹っ掛ける (*fukkakeru*) / overcharge — 請求する (*seikyusuru*) / request

The Japanese WordNet is being used for various purposes including its usage in Weblio's English-Japanese and Japanese-English dictionary[*4]. It can also be used for search query expansion or paraphrase recognition like the case of Case Base for Basic Semantic Relations. "Case Base for Basic Semantic Relations" contains a large number of proper nouns and technical terms as described in Subsection **3.1**, while the Japanese WordNet mainly targets on collecting common words, thus complementing each other.

4 Databases of Semantic Relations between Predicates

4.1 Verb Entailment Database

The database contains 121,508 pairs of

*4 http://ejje.weblio.jp/

verbs: 52,689 pairs of verbs that have an entailment relation and 68,819 pairs of verbs that do not have such relation. A verb pair that has an entailment relation is a pair of verbs where the verb1 cannot be done unless the verb2 is, or has been, done. For example, the acts of "playing in the starting lineup", "microwaving", "sneering", "getting drunk" and "borrowing" entail "starting a game", "warming", "laughing", "drinking" and "lending" respectively.

Information about entailment relations plays an important role in natural language information processing systems. For example, when a question answering system receives the question, "Who started the game between the Giants and the Tigers last night?", the system is required to know that the act of "playing in the starting lineup" entails the act of "starting the game" since the system needs to retrieve the answer by identifying sentences whose surface information is largely different from the information given by the question, such as "Kubo played in the Giants' starting lineup in last night's game against the Tigers", out of a huge amount of documents like Web documents.

The negative instances (verb pairs that do not have an entailment relation) and the positive instances (those that have an entailment relation) in the database can be combined for being used as an input data for machine learning. They can be a set of training data for a machine to learn a model for judging whether an entailment relation exists between two verbs.

The negative and positive instances are classified into 4 subclasses. Each subclass and their examples will be explained in the following subsections. All the negative and positive instances were automatically acquired by using the method proposed by Hashimoto et al. [17] [18] and manually inspected. In the examples below, verbs positioned left of an arrow represent what entails the other and will be called "verb 1", and verbs positioned right of an arrow is what is entailed and will be called "verb 2".

4.1.1 Positive instances

The total number of positive instance pairs is 52,689 and the total numbers of unique verbs 1 and verbs 2 are 36,058 and 8,771 respectively.

Synonymic or hypernym/hyponym pairs that have an entailment relation: The pairs categorized in this group are verb pairs where the verb 1 and the verb 2 have entailment and either of synonymic or hypernym/hyponym relations. Synonymic or hypernym/hyponym pairs that have an inclusive relation in their surface form and are related by the entailment relationship are not listed here but will be presented next. The total number of pairs is 33,802 and the total numbers of unique verbs 1 and verbs 2 are 18,128 and 7,650 respectively. Their examples are given below.

- ○ 挑戦する (*chosensuru*) / try → チャレンジする (*charenjisuru*) / challenge
- ○ チンする (*chinsuru*) / microwave → 加熱する (*kanetsusuru*) / warm
- ○ 同乗する (*dojosuru*) / ride together → 乗る (*noru*) / ride
- ○ 組み立てる (*kumitateru*) / assemble → 作る (*tsukuru*) / make
- ○ 代用する (*daiyosuru*) / substitute → 使う (*tsukau*) / use

Synonymic or hypernym/hyponym pairs that have an inclusive relation in their surface form and are related by entailment relationship: The pairs categorized in this group are synonymic or hypernym/hyponym pairs that have an inclusive relation in their surface form and are related by the entailment relationship. The total number of pairs is 15,599 and the total numbers of unique verbs 1 and verbs 2 are 15,367 and 2,440 respectively. Their examples are given below.

- ○ あざ笑う (*azawarau*) / sneer → 笑う (*warau*) / laugh
- ○ セリーグ優勝する (*seriguyushosuru*) / win the Central League pennant → リーグ優勝する (*riguyushosuru*) / win the league pennant
- ○ 流れ出る (*nagarederu*) / flow out → 出る

(*deru*) / go out
- ○ そそり立つ (*sosoritatsu*) / tower → 立つ (*tatsu*) / stand
- ○ 一部免除する (*ichibu menjosuru*) / partially exempt → 免除する (*menjosuru*) / exempt

Presuppositive relation: A verb pair that has a presuppositive relation is a pair where the verb 2 is the presupposition of the verb 1. In the previously described 2 types of entailment relations, the situations or actions represented by the verb 1 and 2 co-occur, while in a presuppositive relation, the situation or action represented by the verb 2 precedes that of the verb 1. The total number of pairs is 2,846 and the total numbers of unique verbs 1 and verbs 2 are 2,227 and 711 respectively. Their examples are given below.

- ○ 酔っぱらう (*yopparau*) / get drunk → 飲む (*nomu*) / drink
- ○ 稲刈する (*inekarisuru*) / reap rice → 田植する (*tauesuru*) / plant rice
- ○ 乗捨てる (*norisuteru*) / get off → 乗る (*noru*) / get on
- ○ 離職する (*rishokusuru*) / leave one's job → 働く (*hataraku*) / work
- ○ 首席卒業する (*shusekisotsugyosuru*) / graduate as the top student → 学ぶ (*manabu*) / study

Action/reaction relation: A pair of verbs that have an action/reaction relation is a pair where one verb represents an action and the other represents the reaction to it. The verbs 1 and 2 have different agents while in the previously described 3 types of relations, all verbs have the same agents. The total number of pairs is 442 and the total numbers of unique verbs 1 and verbs 2 are 336 and 328 respectively. Their examples are given below.

- ○ 借りる (*kariru*) / borrow → 貸す (*kasu*) / lend
- ○ 受取る (*uketoru*) / receive → 手渡す (*tewatasu*) / hand out
- ○ 教える (*oshieru*) / teach → 学ぶ (*manabu*) / learn
- ○ 売る (*uru*) / sell → 買う (*kau*) / buy
- ○ 預ける (*azukeru*) / entrust → 預かる (*azukaru*) / keep

4.1.2 Negative instances

The total number of negative instance pairs is 68,819 and the total numbers of unique verbs 1 and verbs 2 are 14,658 and 7,077 respectively.

Pairs of associated verbs with no entailment, antonymic or implicational relations: These are pairs of verbs that do not have either of entailment, antonymic or implicational relations but somehow, are associated with each other. Antonymic and implicational relations will be described later. Note that the pairs presented here do not include "pairs of associated verbs that have an inclusive relation in their surface form but do not have either of entailment, antonymic or implicational relations". Those pairs will be presented next. The total number of pairs is 68,306 and the total numbers of unique verbs 1 and verbs 2 are 14,168 and 7,006 respectively. Their examples are given below.

- ○ 通勤する (*tsukinsuru*) / commute → 走る (*hashiru*) / run
- ○ 読書する (*dokushosuru*) / read a book → 寛ぐ (*kutsurogu*) / get relaxed
- ○ ブログ巡りする (*burogumegurisuru*) / surf the Internet visiting blogs → 休む (*yasumu*) / take a break
- ○ 農業体験する (*nogyotaikensuru*) / experience agricultural work → 住む (*sumu*) / live
- ○ 押し黙る (*oshidamaru*) / keep silent → 俯く (*utsumuku*) / drop one's eyes or head

Pairs of associated verbs that have an inclusive relation in their surface form but do not have either of entailment, antonymic or implicational relations: Among the pairs of associated verbs that do not have either of entailment, antonymic or implicational relations, the pairs where the surface form of the verb 2 is included in that of the verb 1 are classified here. The total number of pairs is 294 and the total numbers of unique verbs 1 and verbs 2 are 290 and 101 respectively. Their examples are given below.

○ 冴渡る (*saewataru*) / become clear → 渡る (*wataru*) / pass
○ 準優勝する (*junyushosuru*) / finish second → 優勝する (*yushosuru*) / finish first
○ 怒り出す (*okoridasu*) / get angry → 出す (*dasu*) / take out
○ 歌い上げる (*utaiageru*) / sing in a loud voice → 上げる (*ageru*) / raise
○ 解毒する (*gedokusuru*) / detoxify → 毒する (*dokusuru*) / corrupt

Antonymic relation: These are the pairs of verbs that have an antonymic relation. The total number of pairs is 51 and the total numbers of unique verbs 1 and verbs 2 are 46 and 42 respectively. Their examples are given below.

○ 閉める (*shimeru*) / close → 開ける (*akeru*) / open
○ 反比例する (*hanpireisuru*) / be in inverse proportion → 比例する (*hireisuru*) / be in proportion
○ 失う (*ushinau*) / lose → 得る (*eru*) / obtain
○ 下げる (*sageru*) / lower → 上げる (*ageru*) / raise
○ 飛び去る (*tobisaru*) / fly away → 飛来する (*hiraisuru*) / come flying

Implicational relation: Among the pairs of verbs that cannot be exactly said to have an entailment relation, the pairs where the situation or action represented by the verb 1 can be highly possibly accompanied by the situation or action represented by the verb 2. The total number of pairs is 168 and the total numbers of unique verbs 1 and verbs 2 are 154 and 121 respectively. Their examples are given below.

○ 紅葉する (*koyosuru*) /(of leaves) turn red → 落葉する (*rakuyosuru*) /(of leaves) fall
○ 深煎りする (*fukairisuru*) / roast dark → 挽く (*hiku*) / grind
○ 入会希望する (*nyukaikibosuru*) / hope to be a member → 入会する (*nyukaisuru*) / become a member
○ 印刷プレビューする (*insatsupurebyusuru*) / preview the print → 印刷する (*insatsusuru*) / print
○ 受験する (*jyukensuru*) / take an entrance exam → 進学する (*shingakusuru*) / get enrolled

4.2 Predicate Phrase Entailment Database

This database has not been yet published but is planned to be published shortly with almost 600,000 pairs. It is a collection of pairs of predicate phrases that have an entailment relation (positive instances) and that do not have an entailment relation (negative instances). Verb Entailment Database described above handles entailment relations between words while Predicate Phrase Entailment Database handles those between phrases. The following are their examples.

○ すべての債務を免除される → 債務の支払責任を免除してもらう
(*subete no saimu wo menjosareru → saimu no shiharaisekinin wo menjoshitemorau*)
get exempted from all the debts → get rid of the liability for payment
○ 地球全体の平均気温が上昇する → 地球規模で気温が上昇していく
(*chikyuzentai no heikinkion ga joshosuru → chikyukibo de kion ga joshoshiteiku*)
the average temperature of the earth rises → the temperature rises on a global scale
○ 粉塵を吸入する → ほこりを吸い込む
(*funjin wo kyunyusuru → hokori wo suikomu*)
inhale dust → breathe in dust
○ インシュリンの量が不足する → インスリンの作用が弱くなる
(*inshurin no ryo ga fusokusuru → insurin no sayo ga yowakunaru*)
do not have enough insulin → insulin become less effective
○ 現金でトレードする → お金で取引する
(*genkin de toredosuru → okane de torihikisuru*)
trade in cash → trade in money

Like verb entailment relations, information about entailment relations between predicate phrases also plays an important role in natural language information processing systems. For example, when a question answering system receives the question “What causes cellular aging?”, the system is required to know that

the act of "causing cellular oxidation" entails the act of "causing cellular aging" since the system needs to retrieve the answer by identifying sentences whose surface information is largely different from the information given by the question, such as "DNA damage can cause cellular oxidation", out of a huge amount of documents like Web documents.

The phrases in the database can be classified into two groups of positive instances and negative instances like Verb Entailment Database. The negative and positive instances in the database can be combined for being used as an input data for machine learning. They can be a set of training data for a machine to learn a model for judging whether an entailment relation exists between two predicate phrases.

All the negative and positive instances were automatically acquired from definition sentences in Web documents by using the method proposed by Hashimoto et al. [19] [20]. Part of the acquired phrases will be manually inspected before being released and the rest will be released without further inspection.

The phrases in the database are classified based on their semantic compositionality into two groups of the "perfectly compositional phrase pairs" and the "partially compositional phrase pairs". In the former pair, every content word in one phrase has its counterpart, i.e. synonym or near-synonym, in other phrase. For example, a pair of phrases "合鴨を水田に放す (*aigamo wo suiden ni hanasu*) / release aigamo ducks into a rice paddy → 田にアイガモを放す (*ta ni aigamo wo hanasu*) / into a paddy field, release aigamo ducks" is classified as a perfectly compositional pair of phrases since all the content words in one phrase have their synonyms in the other. On the other hand, if at least one content word in one phrase of a pair does not have its synonym or near-synonym in other phrase, that pair is classified as a partially compositional pair. For example, a pair "地震の揺れを建物に伝わりにくくする (*jishin no yure wo tatemono ni tsutawarinikuku suru*) / prevent transmission of seismic vibration to building structures → 建物自体の揺れを小さくする (*tatemono jitai no yure wo chisaku suru*) / make vibration in building structures smaller" is classified as a "partially compositional phrase pair" since the content words "地震 (*jishin*) / seismic", "伝わる (*tsutawaru*) / transmission" and "小さい (*chisai*) / smaller" do not have their synonyms or near-synonyms in their partner phrase.

We suppose that phrase pairs that are highly semantically compositional can be more easily automatically identified to be related by the entailment relationship than those that are less semantically compositional. This means that the classification of predicate phrases in the database actually reflects the degree of difficulty in identifying entailment relations.

Below are some examples of "perfectly compositional phrase pairs" and "partially compositional phrase pairs".

○ Perfectly compositional phrase pairs
- 生薬をいくつも組み合わせる → いくつもの生薬を組み合わせる
  (*shoyaku wo ikutsu mo kumiawaseru → ikutsu mo no shoyaku wo kumiawaseru*)
  mix various herbal remedies → mix numbers of herbal remedies
- エネルギーが光になる → エネルギーが光となる
  (*enerugi ga hikari ni naru → enerugi ga hikari to naru*)
  energy becomes light → energy becomes light
- 個人情報の取り扱い方法を定める → 個人情報の取扱い方法を定める
  (*kojinjoho no toriastukaihoho wo sadameru → kojinjoho no toriastukaihoho wo sadameru*)
  fix personal information handling policies → fix personal information handling policies
- インターネット上のマナーのことだ → ネットワーク上のエチケットのことだ
  (*intanettojo no mana no kotoda → nettowakujo no echiketto no kotoda*)
  it denotes the manners on the Internet → it denotes the etiquette on the Internet
- 介護サービス計画を作成する → ケア

プランを作成する
(*kaigosabisukeikaku wo sakuseisuru → keapuran wo sakuseisuru*)
make a care service plan → make a care plan
— 文科省が推進している → 文部科学省が推進する
(*monkasho ga suishinshiteiru → monbukagakusho ga suishinsuru*)
be being promoted by the MEXT → be promoted by the Ministry of Education, Culture, Sports, Science & Technology
— アメリカで考案される → 米国で生まれる
(*amerika de koansareru → beikoku de umareru*)
be created in America → be born in the U.S.
— コンピューターに記憶させておく → PC に保存しておく
(*konpyuta ni kiokusaseteoku → pishi ni hozon shiteoku*)
be stored on a computer → be stored on a PC
— パワーが宿る → 力を秘めている
(*pawa ga yadoru → chikara wo himeteiru*)
have power → have hideen power

○ Partially compositional phrase pairs
— かみ合わせや歯並びを回復する → 噛み合わせを復元する
(*kamiawase ya hanarabi wo kaifukusuru → kamiawase wo fukugensuru*)
restore the occlusion and dentition → reconstruct the occlusion
— 悪性細胞が認められる → がん細胞が発生する
(*akuseisaibo ga mitomerareru → gansaibo ga hasseisuru*)
malignant cells are detected → cancer cells grow
— シワやシミを解消する → しわなどを改善する
(*shiwa ya shimi wo kaishosuru → shiwanado wo kaizensuru*)
get rid of wrinkles and spots → improve wrinkles
— 無線 LAN アクセスポイントを共有する → アクセスポイントを公開する
(*musenran akusesu pointo wo kyoyusuru → akusesu pointo wo kokaisuru*)
share a wireless access point → make a wireless access point public
— オートバイで旅行する → バイクで走る
(*otobai de ryokosuru → baiku de hashiru*)
travel on a motorcycle → ride a motorcycle
— 会員間でクルマを共同利用する → クルマを複数の人間で共同利用する
(*kaiinkan de kuruma wo kyodoriyosuru → kuruma wo fukusu no ningen de kyodoriyo suru*)
share one care among the members → share one car among several people
— 電気エネルギーを使用している → エネルギーを電気でまかなう
(*denkienerugi wo shiyoshiteiru → enerugi wo denki de makanau*)
use electrical energy → resort to electricity for power generation
— 情報共有を図る → コミュニケーションを取る
(*johokyoyu wo hakaru → komyunikeshon wo toru*)
try to share information → communicate with each other
— もずくやコンブに含まれている → 海藻類の中に含まれる
(*mozuku ya konbu ni fukumareteiru → kaisorui no naka ni fukumareru*)
be contained in mozuku seaweed or kelp → be contained in seaweed
— コレステロールや中性脂肪の割合が高い → 脂質の値が高い
(*koresuteroru ya chuseihibo no wariai ga takai → shishitsu no atai ga takai*)
cholesterol or neutral fat ratios become high → a fat value becomes high

4.3 Excitatory/Inhibitory Template Database

Excitatory/Inhibitory Template Database is a language resource that lists what we call Excitatory/inhibitory templates. It is planned

to be released around the end of this fiscal year with about 10,000 templates in it. Excitation/inhibition is a new semantic orientation that we had proposed in the References [21][22]. In that framework, phrases consisting of "a joshi (a Japanese postposition) + a predicate" (henceforth called "templates") such as "が発生する (*ga* (joshi) + *hasseisuru* (predicate) / occur)" and "を　防　ぐ (*wo* (joshi) + *fusegu* (predicate) / prevent)" are grouped into three categories of "excitatory", "inhibitory" and "neutral".

Excitatory template: Excitatory templates entail that the main function, effect, purpose, role or impact of the referent of the argument (e.g., the subject or the object) is prepared or activated (e.g., "to cause [something]", "to use [something]", "to buy [something]" "to make [something] progress", "to export [something]", "[of something] to increase", "[of something] to become possible").

Inhibitory template: Inhibitory templates entail that the main function, effect, purpose, role or impact of the referent of the argument is deactivated or suppressed (e.g., "to prevent [something]", "to discard [something]", "to remedy [something]", "[of something] to decrease", "[of something] to be disabled").

Neutral template: Neutral templates are neither excitatory nor inhibitory (e.g., "to consider [something]", "to search for [something]", "to be proportional to [something]").

For example, the phrases "cause an earthquake" and "prevent a tsunami" entail the activation of the impact of "an earthquake" and the deactivation of the impact of "a tsunami" respectively.

Excitation/inhibition is different from certain semantic orientations presented in the References [23][24] such as "good/bad". For example, "get improved" and "have the symptoms of something" are both classified as "excitatory" in our framework, but only the former is classified as "good" in the good/bad semantic orientation, or "remedy something" and "be halted" are both "inhibitory" but only the latter is judged "bad".

Excitatory/inhibitory templates can be used for various purposes. We will present their usages in constructing Predicate Phrase Contradiction Database and Predicate Phrase Causality Database in Subsections **4.4** and **4.5** respectively.

Excitatory/Inhibitory Template Database was constructed by first automatically acquiring candidate templates using the methods we had developed [21] [22] and then manually inspecting them. Below are some examples of excitatory/inhibitory templates planned to be listed in the database.

- Examples of excitatory templates (X: an argument)
  - *X wo takameru* / increase X (を高める)
  - *X wo yuhatsusuru* / induce X (を誘発する)
  - *X wo soshikisuru* / form X (を組織する)
  - *X wo okasu* / commit X (を犯す)
  - *X wo seijokasuru* / normalize X (を正常化する)
  - *X wo jutensuru* / fill X (を充填する)
  - *X de niru* / cook with/by/on X (で煮る)
  - *X ga kotosuru* / X rises (が高揚する)
  - *X ga hofuda* / have plenty of X (が豊富だ)
  - *X ni tassuru* / reach X (に達する)
- Examples of inhibitory templates
  - *X wo mahisaseru* / paralyze X (を麻痺させる)
  - *X wo damasu* / deceive X (を騙す)
  - *X wo shikameru* / knit X (を顰める)
  - *X wo hinansuru* / blame X (を非難する)
  - *X wo shizumeru* / calm X (を静める)
  - *X ni sakarau* / defy X (に逆らう)
  - *X ga suitaisuru* / X declines (が衰退する)
  - *X ga dassensuru* / X derails (が脱線する)
  - *X ga morokusnaru* / X becomes weak (が脆くなる)
  - *X de shippaisuru* / fail in X (で失敗する)

4.4 Predicate Phrase Contradiction Database

The database lists pairs of predicate phrases related by the contradiction relationship (positive instances) such as "destroy cancer ⊥ develop cancer" and those not related by the contradiction relationship (negative instances) such as "get cancer ⊥̸ study cancer". The database is planned to be released around the end of FY 2012 and will contain about a million pairs of predicate phrases including both positive and negative instances. All the predicate phrases in the database consist of three elements (a noun, a joshi (a Japanese postposition) and a predicate) and each element consists of one Japanese word. For example, a phrase "癌を破壊する (*gan wo hakaisuru*) / destroy cancer" consists of "癌(*gan*) / cancer", "を(*wo*, joshi)" and "破壊する (*hakaisuru*) / destroy". For all "a joshi and a predicate" parts, we have used excitatory and inhibitory templates presented in Subsection **4.3**.

A contradiction phrase pair is a pair of predicate phrases that the situation or action represented by one predicate phrase cannot co-occur or coexist with that of the other. Besides these pairs, we added pairs related by what we call quasi-contradiction relationship as a type of positive instances. The requirements to be a predicate phrase pair that have a quasi-contradiction relation are as below.

1. The situation or action represented by one predicate phrase in a pair can co-occur or coexist with that of the other.
2. However, those situations or actions cannot co-occur or coexist (i.e. contradict each other) when the tendency of what one phrase or both phrases represent become extreme.

One example of quasi-contradiction relation pairs is "have tension ⊥ lessen tension". To lessen tension does not always mean its complete disappearance. In other words, one may still have tension. Those two situations can coexist. Therefore, the phrases cannot be said to perfectly contradict each other. However, when the states of "having tension" and "lessening tension" both become extreme, they cannot coexist and thus are judged to have a contradiction relation. In other words, the states of feeling extreme tension and lessening tension completely (i.e. having no tension) have a contradiction relation. Therefore, the pair "have tension ⊥ lessen tension" is classified as a predicate phrase pair having what we call quasi-contradiction relation.

Below are examples of contradiction and quasi-contradiction relation predicate phrase pairs

- Contradiction Relation
  - アンバランスを是正する⊥アンバランスを生じさせる
    (*anbaransu wo zeseisuru ⊥ anbaransu wo shojisaseru*)
    correct an imbalance ⊥ generate an imbalance
  - 円安が止まる⊥円安が進行する
    (*enyasu ga tomaru ⊥ enyasu ga shinkosuru*)
    appreciation of the yen stops ⊥ appreciation of the yen continues
  - 騒音がひどくなる⊥騒音は減少する
    (*soon ga hidokunaru ⊥ soon wa genshosuru*)
    the noise has gotten worse ⊥ the noise has been reduced
  - 酸味がます⊥酸味が消える
    (*sanmi ga masu ⊥ sanmi ga kieru*)
    become more sour ⊥ lose its sour taste
  - 原発をなくす⊥原発を増やす
    (*genpatsu wo nakusu ⊥ genpatsu wo fuyasu*)
    abolish nuclear power plants ⊥ increase the number of nuclear power plants
  - ユーロが下落する⊥ユーロが強くなる
    (*yuro ga gerakusuru ⊥ yuro ga tsuyokunaru*)
    the euro sags ⊥ the euro becomes stronger
  - ウイルスが死滅する⊥ウイルスが活性化する
    (*uirusu ga shimetsusuru ⊥ uirusu ga kasseikasuru*)
    the virus is killed ⊥ the virus is activated

○ Quasi-contradiction Relation
— 痛みが発症する⊥痛みを減らす
(*itami ga hasshosuru* ⊥ *itami wo herasu*)
grow pain ⊥ reduce pain
— アクセスが生ずる⊥アクセスを抑制する
(*akusesu ga shozuru* ⊥ *akusesu wo yokuseisuru*)
have access ⊥ suppress access
— 放射能が放出される⊥放射能が減る
(*hoshano ga hoshutsusareru* ⊥ *hoshano ga heru*)
radioactive substances are emitted ⊥ radioactive substances are reduced
— シェアを有する⊥シェアが低下する
(*shea wo yusuru* ⊥ *shea ga teikasuru*)
have a share ⊥ share declines

Information about contradiction relations between predicate phrases can play an important role in natural language information processing systems. One example is their usage in Web information analysis systems including WISDOM developed by NICT[*5]. Web information analysis systems are required to automatically identify contradictions between informations given by Web documents so that the system can provide its users with opposing opinions or information. For example, when a system receives the question "What will be the impact on the environment if we halt nuclear power plant operations?", the system may find contradicting descriptions in different Web documents. One document may write "We can protect the environment by halting nuclear power plant operations because they can contaminate our environment by emitting nuclear substances" and the other may write "Halting nuclear power generation may increase the ratio of thermal power generation and CO_2 emission, leading to the deteriorated environment." The system is required to automatically identify contradicting points in these two documents and sum up opposing opinions to provide the user with appropriate information.

The phrases in the database can be classified into two groups of positive instances and negative instances like Verb Entailment Database and Predicate Phrase Entailment Database. The negative and positive instances in the database can be combined for being used as an input data for machine learning. They can be a set of training data for a machine to learn a model for judging whether a contradiction relation exists between two predicate phrases.

All the positive and negative instances were prepared from the results automatically acquired by using the method proposed by Hashimoto et al. [21] [22]. The precision rate for the automatic acquisition was 70% among the million top-scoring pairs. The method used for detecting contradiction relations utilized the excitatory/inhibitory templates automatically acquired by using the same method by Hashimoto et al. [21] [22]. To be concrete, the contradiction relation phrase pair "destroy cancer ⊥ develop cancer" can be obtained by combining a noun (cancer) and a pair of excitatory/inhibitory templates that have opposite orientations "destroy (something)" and "develop (something)" (the former template is inhibitory and the latter is excitatory).

4.5 Predicate Phrase Causality Database

The database lists pairs of predicate phrases related by the causal relationship (positive instances) such as "smoke cigarettes ⇒ have lung cancer" and those not related by the causal relationship (negative instances) such as "smoke cigarettes ⇒ go to the company". The database is planned to be released around the end of FY 2012 and will contain about a million pairs of predicate phrases including both positive and negative instances. All the predicate phrases in the database consist of three elements (a noun, a joshi (a Japanese postposition) and a predicate) and each element consists of one Japanese word. For example, a phrase "肺癌になる (*haigan ni naru*) / have lung cancer" consists of "肺癌 (*haigan*) / lung cancer", "に(*ni*, joshi)" and "なる (*naru*) / have". As in

*5 http://wisdom-nict.jp/

"Predicate Phrase Contradiction Database" presented in Subsection **4.4**, we have used excitatory and inhibitory templates presented in Subsection **4.3** for all "a joshi and a predicate" parts.

Below are examples of predicate phrase pairs planned to be listed in the database

- 基礎代謝を高める⇒脂肪燃焼力を高める
  (*kisotaisha wo takameru* ⇒ *shibonenshoryoku wo takameru*)
  increase the basal metabolism rate ⇒ increase the fat burn ability
- 学習意欲を高める⇒自己学習を促進する
  (*gakushuiyoku wo takameru* ⇒ *jikogakushu wo sokushinsuru*)
  enhance motivation to learn ⇒ promote self-learning
- 輸出が増える⇒ GDP が増加する
  (*yushutsu ga fueru* ⇒ *GDP ga zokasuru*)
  have increased import ⇒ have a higher GDP
- 血行を促進する⇒新陳代謝を助ける
  (*kekko wo sokushinsuru* ⇒ *shinchintaisha wo tasukeru*)
  facilitate the flow of blood ⇒ contribute to better basal metabolism
- 視界が良くなる⇒作業効率が向上する
  (*shikai ga yokunaru* ⇒ *sagyokoritsu ga kojosuru*)
  have a better view ⇒ improve operational efficiency
- 大地震が発生する⇒メルトダウンを起こす
  (*daijishin ga hasseisuru* ⇒ *merutodaun wo okosu*)
  have a catastrophic earthquake ⇒ have a nuclear meltdown
- 熱効率が良い⇒暖房効果を高める
  (*netsukoritsu ga yoi* ⇒ *danbokoka wo takameru*)
  have higher thermal efficiency ⇒ improve effects of heating
- インフレを起こす⇒円安が進行する
  (*infure wo okosu* ⇒ *enyasu ga shinkosuru*)
  cause inflation ⇒ promote yen's appreciation
- 体力が落ちる⇒免疫力が下がる
  (*tairyoku ga ochiru* ⇒ *menekiryoku ga sagaru*)
  lose physical strength ⇒ have reduced immune strength
- 国債先物急落を受ける⇒金利が上昇する
  (*kokusaisakimonokyuraku wo ukeru* ⇒ *kinri ga joshosuru*)
  see a sharp drop in government bond futures prices ⇒ see an interest rate hike

A pair of predicate phrases that have a causal relation in the database is a pair where the possibility of the occurrence or existence of the event, act or state represented by the phrase positioned right becomes higher when the event, act or state represented by the phrase positioned left occurs or exists compared with the case of no such occurrence or existence (the event, act or state represented by the left-side phrase should occur almost simultaneously with or precede that of the right-side phrase). This means that causal relations in this database do not always provide information that the occurrence or existence of the event, act or state represented by the left-side phrase always means the occurrence or existence of such situations represented by the right-side phrase. For example, although the phrase pair "have a catastrophic earthquake ⇒ have a nuclear meltdown" is listed as a causal pair in the database, this does not mean that a catastrophic earthquake always leads to a nuclear meltdown. The pair was judged to have a causal relation just because the possibility of having a meltdown becomes higher when a catastrophic earthquake happens compared with the case of no occurrence of such an earthquake.

Furthermore, we have established two standards for judging whether a phrase pair should be listed as a causal pair in our database. We call them the generality standard and the standard for unverified cases. The former states that a phrase pair that represents causality that is too exceptional or lacks generality should not be included in the database even if the phrases are used in such way that they have a causal relation in the documents they had been extracted from. For example, if there is a sentence "Let's have vegetarian dishes for the New Year's party because Mr. Ichikawa

will be joining us" in a corpus, "Mr. Ichikawa will be joining us ⇒ have vegetarian dishes" should not be listed as a causal pair since it is too exceptional and lacks generality. The standard for unverified cases states that a phrase pair that represents causality that has not been scientifically verified should be judged as a causal pair if you find at least one evidence to support that causal relation in Web documents. For example, a phrase pair "drink black oolong tea ⇒ suppress fat absorption" should be judged as a causal pair if there is a descriptions like "I heard black oolong tea suppresses fat absorption" in Web documents.

Thus, the users of the database should note that the phrase pairs listed in this database do not always provide accurate information of causal relations. The pairs in the database had been manually inspected, but still, their inspection and judgment were based on the knowledge provided by Web documents and this does not necessarily mean that causal relations that were judged to be reasonably causal based on such knowledge are always and absolutely true.

All the causal and non-causal pairs were acquired by using two methods for automatically identifying causal relations presented in the References [21][22]. One is the method to automatically extract causal pairs in Web documents (henceforth called the method for extracting causality) and the other is the method to automatically generate causal pairs that have a highly possible causal relation without verification by Web documents (henceforth called the method for generating causality hypothesis). The method for extracting causality extracts causal pairs by identifying two combinations of an excitatory/inhibitory template and a noun co-occurring and being connected by a resultive conjunction in a sentence on a Web document. For example, a sentence "犯罪が増加すると不安が高まる / The number of criminal cases increases and people's anxiety gets heightened" has two combinations of "が増加する / increases" (excitatory/inhibitory template) and "犯罪 / the number of criminal cases" (noun) and "が 高 ま る / gets heightened" and "不 安 / people's anxiety" (noun), and they are connected by the resultive conjunction "と / and" in one sentence, therefore the phrase are extracted as a causal pairs "犯罪が増加する⇒不安が高まる / the number of criminal cases increases ⇒ people's anxiety gets heightened". The precision rate for the automatic extraction was 70% among the 500,000 top-scoring pairs. As for the method for generating causality hypothesis, it automatically generates hypothetically causal relations (e.g. "decrease the number of criminal cases ⇒ anxiety disappears") by replacing one phrase in an automatically extracted pair (e.g. "the number of criminal cases increases ⇒ people's anxiety gets heightened) with a contradictory phrase (e.g. "the number of criminal cases increases ⊥ decrease the number of criminal cases" and "people's anxiety gets heightened ⊥ anxiety disappears", generating "decrease the number of criminal cases ⇒ anxiety disappears". For details, see Subsection **4.4**). Note that if two phrases that have a hypothetically causal relation are found within a sentence on a Web document, those phrases are not judged to be a hypothetically causal pair. This means that the database includes not only causal pairs found on the Web but causal pairs that may have a highly possible causal relation despite the fact that their relationship is not explicitly stated in a Web document. The precision rate for the automatic generation was 57% among the million top-scoring pairs. Below are English translations of some examples of hypothetically causal pairs planned to be included in the database. Written between brackets are causal relations originally found on the Web and used as the base of hypothesis generation.

- ストレスが減少する⇒不眠が改善される
  （ストレスが増加する⇒不眠が続く）
  sutoresu ga genshosuru ⇒ fumin ga kaizensareru
  (*sutoresu ga zokasuru ⇒ fumin ga tsuzuku*)
  reduce stress ⇒ get rid of sleeplessness
  (have increased stress ⇒ sleeplessness continues)
- デフレを阻止する⇒税収が増加する

（デフレが進む⇒税収が減る）
defure wo soshisuru ⇒ zeishu ga zokasuru
(*defure ga susumu ⇒ zeishu ga heru*)
avoid deflation ⇒ increase tax revenue
(accelerate deflation ⇒ have decreased tax revenue)

○ 楽しみが増大する⇒ストレスが減少する
（楽しみが減る⇒ストレスが高まる）
tanoshimi ga zodaisuru ⇒ sutoresu ga genshosuru
(*tanoshimi ga heru ⇒ sutoresu ga takamaru*)
have greater hopes ⇒ have less stress
(have less hopes ⇒ heighten stress)

○ 犯罪を減らす⇒不安が無くなる
（犯罪が増加する⇒不安が高まる）
hanzai wo herasu ⇒ fuan ga nakunaru
(*hanzai ga zokasuru ⇒ fuan ga takamaru*)
decrease the number of criminal cases ⇒ anxiety disappears
(the number of criminal cases increases ⇒ anxiety gets heightened)

○ 塩素を減らす⇒バクテリアは増殖する
（塩素を発生させる⇒バクテリアを死滅させる）
enso wo herasu ⇒ bakuteria wa zoshokusuru
(*enso wo hasseisaseru ⇒ bakuteria wo shimetsusaseru*)
reduce the amount of chroline ⇒ bacteria multiply
(generate chroline ⇒ kill bacteria)

○ 需要が拡大する⇒失業を減少させる
（需要が減る⇒失業が増える）
juyo ga kakudaisuru ⇒ shitsugyo wo genshosaseru
(*juyo ga heru ⇒ shitsugyo ga fueru*)
have a greater demand ⇒ lower the unemployment rate
(have a smaller demand ⇒ have a higher unemployment rate)

○ 疲れを軽減する⇒免疫を増強する
（疲れがたまる⇒免疫が弱まる）
tsukare wo keigensuru ⇒ meneki wo zokyosuru
(*tsukare ga tamaru ⇒ meneki ga yowamaru*)
alleviate fatigue ⇒ boost the immune system
(accumulate fatigue ⇒ have a weaker immune system)

○ 調子があがる⇒トラブルを防げる
（調子が悪くなる⇒トラブルが起きる）
choshi ga agaru ⇒ toraburu wo fusegeru
(*choshi ga warukunaru ⇒ toraburu ga okiru*)
improve ⇒ prevent trouble
(be in a bad condition ⇒ have trouble)

4.6 Database of Japanese Paraphrasing Patterns

In obtaining knowledge from a large scale document data such as Web documents, identification of the sentences that have the same or similar meanings and are interchangeable will enable us to acquire a greater amount of knowledge. “Database of Japanese Paraphrasing Patterns” has been constructed by making use of the syntactic analysis results and contains paraphrasable sentence or phrase patterns for a given sentence or phrase. Paraphrasable sentences like “A has plenty of B” have replaceable nominals (A and B in this case) and a pattern that links the nominals. The database contains such paraphrasing patterns and their score to show their likelihood. Examples of the paraphrasing patterns and scores for “A has plenty of B,” “A stops B” and “A makes B happy” are shown in Tables 8, 9 and 10.

The targets of paraphrasing in “Database of Japanese Paraphrasing Patterns” are those obtained from 50 million Web documents. A paraphrasing pattern consists of nouns A and B that have a certain level of appearance frequency and words situated on the dependency path to connect A and B in a syntax tree. For example, from a sentence “交通事故による経済的な損害に関して / regarding economic loss due to a traffic accident” shown in Fig. 1, we can extract a pattern “A による (due to A)”.

The similarities between patterns are obtained based on the distributions of noun pairs positioned at the slots of variables A and B in a pattern. For details, please see the description about the “SC (Single Class)” method in

Table 8 *Paraphrasing patterns of "A は B が豊富です (**A wa B ga hofudesu**) / A has plenty of B"(5 top-scoring patterns)*

Pattern	Paraphrasing Score
〈A は B が豊富 (*A wa B ga hofu*) / A has plenty of B	0.0549719888
〈A には B が豊富に含まれています (*A ni wa B ga hofu ni fukumare-teimasu*) / A contains a lot of B〉	0.0382925298
〈A は B も豊富です (*A wa B mo ho-fudesu*) / A has plenty of B as well〉	0.0377786173
〈A は B を多く含む (*A wa B wo oku fukumu*) / A contains B a lot〉	0.0336538462
〈A は B も豊富 (*A wa B mo hofu*) / A has plenty of B as well〉	0.0331325301

Table 9 *Paraphrasing patterns of "A は B を防ぐ (**A wa B wo fusegu**) / A stops B" (5 top-scoring patterns)*

Pattern	Paraphrasing Score
〈A が B を防ぐ (*A wa B wo fusegu*) / It is A that prevents B〉	0.0224161276
〈A は B を予防する (*A wa B wo yobosuru*) / A prevents B〉	0.0186121788
〈A で B を防ぐ (*A de B wo fusegu*) / B is prevented by A〉	0.0175963197
〈B を防ぐ A (*B wo fusegu A*) / A that prevents B〉	0.0175141447
〈A は B を防止する (*A wa B wo boshisuru*) / A checks B〉	0.0132786565

Table 10 *Paraphrasing patterns of "A で B を喜ばせる (**A de B wo yorokobaseru**) / A makes B happy" (5 top-scoring patterns)*

Pattern	Paraphrasing Score
〈A を B 様にご提供していきたい (*A wo Bsama ni goteikyoshiteikitai*) / We would like to continue to provide Mr./Ms. B with A〉	0.0430107527
〈B 様に A を提供して参りました (*Bsama ni A wo teikyoshitemairi-mashita*) / We have been providing Mr./Ms. B with A〉	0.0337078652
〈A を B 様に提供し続けること (*A wo Bsama ni teikyoshitsuzukeru koto*) / Keeping providing Mr./Ms. B with A〉	0.0337078652
〈B 様に A を提供出来るように (*Bsama ni A wo teikyodekiru yo ni*) / In order for us to provide Mr./Ms. B with A〉	0.0337078652
〈B 様に A を提供出来るよう (*Bsama ni A wo teikyodekiru yo*) / In order for us to provide Mr./Ms. B with A〉	0.0333333333

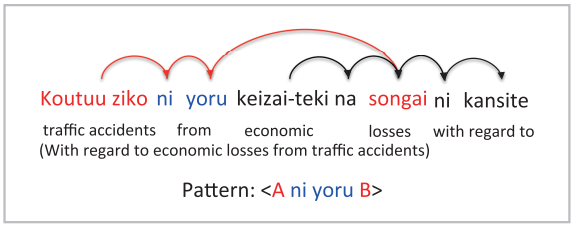

Fig.1 *The pattern extraction out of analysis results of dependency structures*

the reference [6]. Since it is a method of automatic acquisition based on unsupervised learning, the paraphrasing patterns in the database are not always accurate.

In connection with the database, a database of entailment relations between phrase patterns are now being constructed by using the results automatically acquired by the supervised learning-based method proposed by Kloetzer et al. [25]. The precision rate for automatic acquisition was 70% among the 10 million top-scoring pairs. Below are examples of entailment relations between paraphrasing patterns acquired by using the method proposed by Kloetzer et al.

- ○ A を生み出す B → A を作る B
  (*A wo umidasu B → A wo tsukuru B*)
  B that creates A → B that makes A
- ○ A に出向く B → A に行く B
  (*A ni demuku B → A ni iku B*)
  B that visits A → B that goes to A
- ○ A に上程されていた B → A に B を提出する
  (*A ni joteisareteita B → A ni B wo teishu-tsusuru*)
  B that has been presented to A → submit B to A
- ○ A を B に変更 → A を B にする

(*A wo B ni henko → A wo B ni suru*)
change A to B → make A B

- ○ B に光る A → B に輝く A
  (*B ni hikaru A → B ni kagayaku A*)
  A that is shining on B → A that is glittering on B
- ○ A を乗り換えられる B → A を変更できる B
  (*A wo norikaerareru B → A wo henkodekiru B*)
  B where one can transfer to A → B where one can change A
- ○ B の材料を生かした A → B の素材を使った A
  (*B no zairyo wo ikashita A → B no sozai wo tsukatta A*)
  A that utilizes the ingredients of B → A that uses the materials used for B
- ○ A を担いだ B → A を背負った B
  (*A wo katsuida B → A wo seotta B*)
  B that carry A on its shoulder → B that shoulders A
- ○ A が奉られている B → A を祀る B
  (*A ga matsurareteiru B → A wo matsuru B*)
  B that is dedicated to A → B where A is enshrined
- ○ B を強化する A → B を育てる A
  (*B wo kyokasuru A → B wo sodateru A*)
  A that strengthens B → A that develops B

5 Dependency Database and Corpora

5.1 Japanese Dependency Structure Database and Dependency Structure Database of Japanese Wikipedia Entries

"Japanese Dependency Structure Database" and "Dependency Structure Database of Japanese Wikipedia Entries" contain dependency structures and their frequencies obtained by syntactically analyzing a huge amount of Japanese documents and extracting dependency structures from the syntactic analysis results. Table 11 shows their examples.

"Japanese Dependency Structure Database" contains 4.6 billion dependency structures and their frequencies. The dependency structures were extracted from 6 hundred million Web documents and a dependency structure consists of two bunsetsu (a basic unit of Japanese clause) such as "関サバを食べる / eat sekisaba mackerel, broken down to *sekisaba wo* and *taberu*" and "関サバのお作り / sashimi of sekisaba mackerel, broken down to *sekisaba no* and *otsukuri*".

"Dependency Structure Database of Japanese Wikipedia Entries" contains depen-

Table 11 *Examples of dependency structures and their frequencies in 2 dependency structure databases*

Database	Dependency structure	Frequency
Japanese Dependency Structure	関サバを食べる (*sekisaba wo taberu*) / eat sekisaba mackerel	20 times
Japanese Dependency Structure	関サバのお造り (*sekisaba no otsukuri*) / sashimi of sekisaba mackerel	7 times
Japanese Dependency Structure	野球を観戦する (*yakyu wo kansensuru*) / watch (a) baseball (game)	40 times
Japanese Dependency Structure	野球のボール (*yakyu no boru*) / a ball for playing baseball	20 times
Dependency Structure of Wikipedia Entries	風と共に去りぬを借りる (*kaze to tomo ni sarinu wo kariru*) / borrow Gone with the Wind	12 times
Dependency Structure of Wikipedia Entries	三保の松原の景色 (*miho no matsubara no keshiki*) / the view of Miho no Matsubara	6 times
Dependency Structure of Wikipedia Entries	瞬間湯沸かし器で一酸化炭素中毒事故 (*shunkanyuwakashiki de issankatansochudokujiko*) / carbon monoxide poisoning caused by an instantaneous water heater	8 times
Dependency Structure of Wikipedia Entries	星の王子さまを読む (*hoshi no ojisama wo yomu*) / read The Little Prince	3,643 times

dency structures and their frequencies using the same Web documents as those used in "Japanese Dependency Structure Database". While "Japanese Dependency Structure Database" lists only dependency structures consisting of two bunsetsus, "Dependency Structure Database of Japanese Wikipedia Entries" contains dependency structures of Wikipedia article titles (entries) that consist of two or more bunsetsus (e.g. "三保の松原 (*Miho no Matsubara*, a location name)", "風と共に去りぬ (*Kaze to tomo ni Sarinu*, meaning 'Gone with the Wind')"), thus supplementing what "Japanese Dependency Structure Database" lacks, i.e. dependency structures containing named entities and consisting of more than two bunsetsus.

Both "Japanese Dependency Structure Database" and "Dependency Structure Database of Japanese Wikipedia Entries" are indispensable for many language resources that are compiled based on frequencies of dependency structures such as "Database of Similar Context Terms" (Subsection **3.3**). For example, "Database of Similar Context Terms" includes nouns and noun phrases that represent, for example, animation movie titles, famous composers, celebrated conductors or old-time rock bands. Those named entities had been automatically acquired by using the knowledge in the dependency structure databases, i.e. the dependees of nominals in Web documents, as contexts of their appearance. Table 12 shows the dependees of "関サバ / sekisaba mackerel" and "関アジ / sekiaji horse mackerel", i.e. component parts of the contexts of their appearance in Database of Similar Context Terms. You can see the noun phrases "sekisaba mackerel" and "sekiaji horse mackerel" highly frequently appear in the same context since their dependees shown in the table including "の刺身 / sashimi of …", "の活造り / live sashimi of …", "の干物 / dried …" and "がおいしい / … tastes good" are all considered characteristic to words to donate fish and frequently appear with both "sekisaba mackerel" and "sekiaji horse mackerel".

Table 12 *Dependees of "関サバ (**sekisaba**) / sekisaba mackerel" and "関アジ (**sekiaji**) / sekiaji horse mackerel" and their appearance frequencies*

Dependee	"関サバ / Sekisaba"	"関アジ / Sekiaji"
の刺身 (*no sashimi*) / sashimi of …	106 times	92 times
の活造り (*no tsukuri*) / live sashimi of …	12 times	11 times
の干物 (*no himono*) / dried …	15 times	10 times
を仕入れる (*wo shiireru*) / stock …	4 times	4 times
を使う (*wo tsukau*) / use …	10 times	14 times
を堪能 (*wo tanno*) / enjoy …	4 times	6 times
がおいしい (*ga oishii*) / … tastes good	25 times	10 times
を食する (*wo shokusuru*) / eat …	2 times	7 times
は有名だ (*wa yumeida*) / … is famous	9 times	14 times
に劣らない (*ni otoranai*) / be as good as …	4 times	10 times

5.2 Kyoto Sightseeing Blogs for Evaluative Information

Recent advances in information media have allowed many people to publicly express their evaluations and opinions on various issues. Accordingly, studies on technologies to extract, organize and sum up various opinions out of a huge amount of documents is actively being conducted. Kyoto Sightseeing Blogs for Evaluative Information was constructed to serve as a training corpus for machine learning, a basis for developing opinion analysis technologies. The database consists of two parts: "Kyoto Sightseeing Blogs" and "Evaluative Information Data on Kyoto Sightseeing Blogs".

"Kyoto Sightseeing Blogs" is a database containing 1,041 Japanese blog articles (480 Japanese characters per article on the average) exclusively in the tourism domain written by 47 authors. The authors had been recruited with the condition that all copyrights were go-

ing to be reserved by National Institute of Information and Communications Technology. They were asked to write articles based on actual Kyoto sightseeing tours. The authors write their articles by accessing our blog site (not open to the public).

"Evaluative Information Data on Kyoto Sightseeing Blogs" contains evaluative information (popularity and opinions) manually extracted from Kyoto Sightseeing Blogs according to certain standards stated in the References [26][27]. Besides popularity and opinions, evaluative information includes the evaluation holders, expressions used in their evaluation and targets of evaluation. Tables 13 and 14 show examples of articles and their evaluative information respectively. For details about annotation, see the Reference [27].

As shown in Table 14, the database contains not only subjective opinions like "It is beautiful" but objective ones such as "It has been listed as a World Heritage Site" if the part is written in such a way that it describes the good or bad points about the place focused in an article.

Traditionally, training corpora for extracting opinions have been constructed from newspaper articles. However, systems trained on such database can hardly give the highly accurate results since many consumer generated media including blog articles are written in informal or colloquial styles and use emoticons. Therefore, construction of organized training data compiled from blog articles like the data presented here is quite important for developing highly accurate technologies to automatically analyze such informal documents as blogs.

6 Tools, Web Services and Searching Systems

6.1 Hyponymy Extraction Tool

Hyponymy Extraction Tool is a tool to extract hyponymy relations between terms (hypernym/hyponym pairs) from Wikipedia dump data based on the method proposed by Sumida et al. [28]. A hyponymy relation is defined as a relation between two terms X and Y satisfying condition "Y is a kind (an instance) of X". In this section, we denote a hyponymy relation for hypernym X and its hyponym Y as "X → Y". Hypernyms and hyponyms obtained by this tool are not only "words" but "compound

Table 13 *Example of blog article*

ID	Title	Content of article
30	Kamigamo Shrine	Decided to stop by Kamigamo Shrine since we were there. The place is listed as a World Heritage Site, I heard. They say it's one of the oldest shrines in Kyoto. Passing under the torii, a kind of symbolic guard frame at the entrance, situated right across the bus stop, I saw a tree-filled green space. There were several cherry trees. The weeping cherry trees were beautifully blooming. …

Table 14 *Examples of evaluative information*

Topic	ID	Extracted sentence	Evaluative expression	Evaluation type	Evaluation holder	Evaluation target	Relation
Kamigamo Shrine	30	The place is listed as a World Heritage Site, I heard.	is listed as a World Heritage Site, I heard	Merit +	[unknown]	[Kamigamo Shrine]	Same
Kamigamo Shrine	30	They say it's one of the oldest shrines in Kyoto.	They say it's one of the oldest shrines in Kyoto	Merit +	[unknown]	[Kamigamo Shrine]	Same
Kamigamo Shrine	30	The weeping cherry trees were beautifully blooming.	The weeping cherry trees were beautifully blooming	Emotion +	[Author]	[Kamigamo Shrine]	Same

nouns" such as "sports event in Shima City".

For the extraction of term pairs that seem to have a hyponymy relation, i.e. hyponymy relation candidates, we used hierarchical structures, definition sentences and category tags in Wikipedia articles as shown Fig. 2.

Hierarchical Structures: Hyponymy relation candidates are extracted from an article title, session title and itemized expressions in hierarchical structures of Wikipedia articles. For example, "cheese → processed cheese" and "cheese → natural cheese" are extracted as candidates from the example Fig.2 (a).

Definition Sentences: The first sentence in Wikipedia articles is considered as a definition sentence of the article. Hyponymy relation candidates are extracted from these definition sentences by using patterns such as "～とは (… is a …)" and "～の一種 (… is a type of …)". For example, "food → cheese" is extracted as a candidate from the example in Fig. 2 (b).

Category Tags: Hyponymy relation candidates are extracted from all the possible pairs of article title and Wikipedia category tag in a Wikipedia article. For example, "fermented food → cheese" is extracted as a candidate from the example in Fig. 2 (c) (pairs of the same terms such as "cheese → cheese" are excluded from candidates).

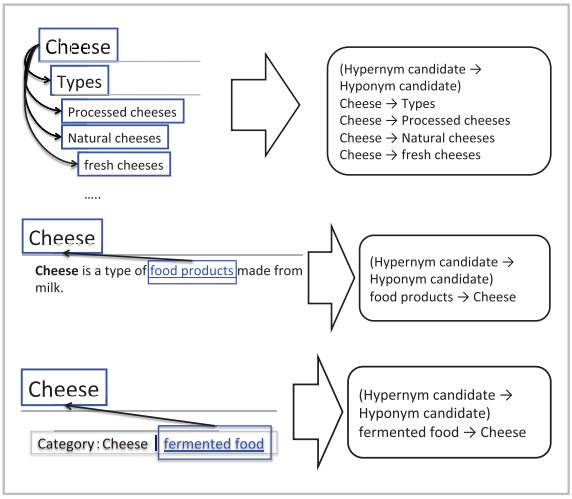

Fig.2 *The extraction of hyponymy candidates out of articles in Wikipedia*

All extracted candidates are judged whether they have a hyponymy relation or not by using SVMs (Support Vector Machines). For training SVMs, we use lexical features such as morpheme and word information of candidates, structural features such as parent and child node in a hierarchical structure of Wikipedia articles from which candidates are extracted, and semantic features derived from Wikipedia infobox. For the details of the algorithm for the acquisition of hyponymy relations, please see the reference by Oh et al. [29] and Sumida et al. [28]

By using this tool, about 7.2 million term pairs having a hyponymy relation were extracted from the May 3, 2012 version of Japanese Wikipedia articles with around 90% precision. Table 15 shows the numbers of hyponymy relations and their unique hypernyms and hyponyms acquired from hierarchical structures, definition sentences and Wikipedia category tags. Table 16 shows examples of acquired hyponymy relations.

6.2 Support Service for Customized Word Set Generation

We have been developing Web services intended to share with general users. Those services have been created by making easily usable the natural language processing technologies and language resources that we have developed and constructed. The Web service presented here allows general users who do

Table 15 *Number of hyponymy relations acquired from the May 3, 2012 version of Japanese Wikipedia*

Source of extraction	# of hyponymy relations	# of unique hypernyms	# of unique hyponyms
Hierarchical structures	5,256,876	153,871	2,670,341
Definition sentences	384,733	40,849	373,580
Category tags	1,766,485	63,876	652,284
Total #	7,217,525	237,593	2,931,627

Table 16 *Examples of acquired hyponymy relations*

Hypernym	Hyponym
仏像 (*butsuzo*) / statue of Buddha	七面大明神像 (*shichimendaimyojinzo*) / statue of Shichimen Daimyojin
ジャズフェスティバル (*jazufesutibaru*) / jazz festival	BAY SIDE JAZZ CHIBA / BAY SIDE JAZZ CHIBA
楽器 (*gakki*) / musical instrument	カンテレ (*kantere*) / kantele
文房具 (*bunbogu*) / stationary	スティックのり (*sutikkunori*) / glue stick
神楽団体 (*kaguradantai*) / kagura troup	川平神楽社中 (*kawahirakagurashachu*) / Kawahira Kagura Troup
プログラミング言語 (*puroguramingugengo*) / programming language	prolog / prolog
戦争映画 (*sensoeiga*) / war film	ハワイ・ミッドウェイ大海空戦 (*hawaimiddoueidaikaikusen*) / Hawai Middouei Daikaikusen
日本映画 (*nihoneiga*) / Japanese film	歌う若大将 (*utau wakadaisho*) / Utau Wakadaisho
AOC ワイン (*AOC wain*) / AOC wine	ラ・グランド・リュー ブルゴーニュ (*ragurandoryu burugonyu*) / La Grande Rue, Bourgogne
ゲーム (*gemu*) / game	ファイナルファンタジー XI (*fainarufantajiXI*) / Final Fantasy XI
テレビ時代劇 (*terebijidaigeki*) / historical TV drama	江戸の渦潮 (*edo no uzu*) / Edo no Uzu (a Japanese samurai TV drama)
放送事業者 (*hosojigyosha*) / broadcasting organization	西日本放送 (*nishinipponhoso*) / Nishinippon Broadcasting Company, Limited
トラス橋 (*torasukyo*) / truss bridge	川島大橋 (*kawashimaohashi*) / Kawashima Bridge
政治制度 (*seijiseido*) / political system	直接民主制 (*chokusetsuminshusei*) / direct democracy
病気 (*byoki*) / disease	セレン欠乏症 (*serenketsubosho*) / selenium deficiency
発電方式 (*hatsudenhoshiki*) / type of power generation	太陽光発電 (*taiyokohatsuden*) / solar power generation
火力発電所 (*karyokuhatsuden*) / thermal power station	ジェネックス水江発電所 (*jenekkusumizuehatsudensho*) / GENEX Mizue power station
羽毛恐竜 (*umokyoryu*) / feathered dinosaurs	シノサウロプテリクス (*shinosauroputerikusu*) / Sinosauropteryx
都市 (*toshi*) / city	バンクーバー (*bankuba*) / Vancouver
市立中学校 (*shiritsuchugakko*) / municipal junior high school	伊佐市立大口南中学校 (*isashiritsuokuchiminamichugakko*) / Isa City Okuchi Minami Junior High School
黄色顔料 (*kiiroganryo*) / yellow pigment	インディアンイエロー (*indianiero*) / Indian yellow
研究所 (*kenkyusho*) / research institute	情報通信研究機構 (*johotsushinkenkyukiko*) / National Institute of Information and Communications Technology

not have special expertise to easily generate groups of words categorized in a certain type of group and word pairs that have a certain semantic relation such as a causal relation. The former service is called "Support Service for Customized Word Set Generation" and the latter is "Semantic Relation Acquisition Service", and both are open to the public. "Support Service for Customized Word Set Generation" is presented in this section and "Semantic Relation Acquisition Service" will be presented in Subsection **6.3**.

"Support Service for Customized Word Set Generation" is a service to allow users to

generate groups of words (word classes) that are semantically similar. Word classes play an important role in various natural language processing systems. For example, they can be used for query expansion in search systems or automatic keyword suggestion for keyword advertising systems.

"Support Service for Customized Word Set Generation" enables efficient semi-automatic generation of a large amount of word classes using Japanese Web documents based on a statistical method. 10 million words on the Web are candidate words to be included in word classes. For the details of the method used for the service, please see the reference [30].

Below are examples of word classes obtained by using the service.

○ "お寺・神社 (*otera/jinja*) / Temple/shrine" class
  — "金閣寺 (*kinkakuji*) / Kinkakuji Temple", "東大寺 (*todaiji*) / Todaiji Temple", "正倉　院 (*shosoin*) / Shosoin Treasure House", "上賀茂神社 (*kamigamojinja*) / Kamigamo Shrine", "銀閣寺 (*ginkakuji*) / Ginkakuji Temple", "三十三間堂 (*sanjusangendo*) / Sanjusangendo Temple", "法隆寺 (*horyuji*) / Horyuji Temple", "平等 院 (*byodoin*) / Byodoin Temple", "清水　寺 (*kiyomizudera*) / Kiyomizudera Temple", "日光東照宮 (*nikkotoshogu*) / Nikko Toshogu Shrine", "善光寺 (*zenkoji*) / Zenkoji Temple", "厳島神社 (*itsukushimajinja*) / Itsukushima Jinja Shrine", "平 安 神 宮 (*heianjingu*) / Heian Jingu Shrine", "中 尊 寺 (*chusonji*) / Chusonji Temple", "出　雲　大　社 (*izumotaisha*) / Izumo Taisha Shrine", "白馬寺 (*hakubaji*) / Hakubaji Temple", "飛　鳥　寺 (*asukadera*) / Asukadera Temple", "明 月 院 (*meigetsuin*) / Meigetsuin Temple", "浅草寺 (*sensoji*) / Sensoji Temple", "三 千 院 (*sanzenin*) / Sanzenin Temple", "薬師寺 (*yakushiji*) / Yakushiji Temple", "南禅寺 (*nanzenji*) / Nanzenji Temple", "室生寺 (*muroji*) / Muroji Temple", "竜安寺 (*ryoanji*) / Ryoanji Temple", "長　谷　寺 (*hasedera*) / Hasedera Temple", "四天王寺 (*shitennoji*) / Shitennoji Temple", "東福寺 (*tofukuji*) / Tofukuji Temple", "唐

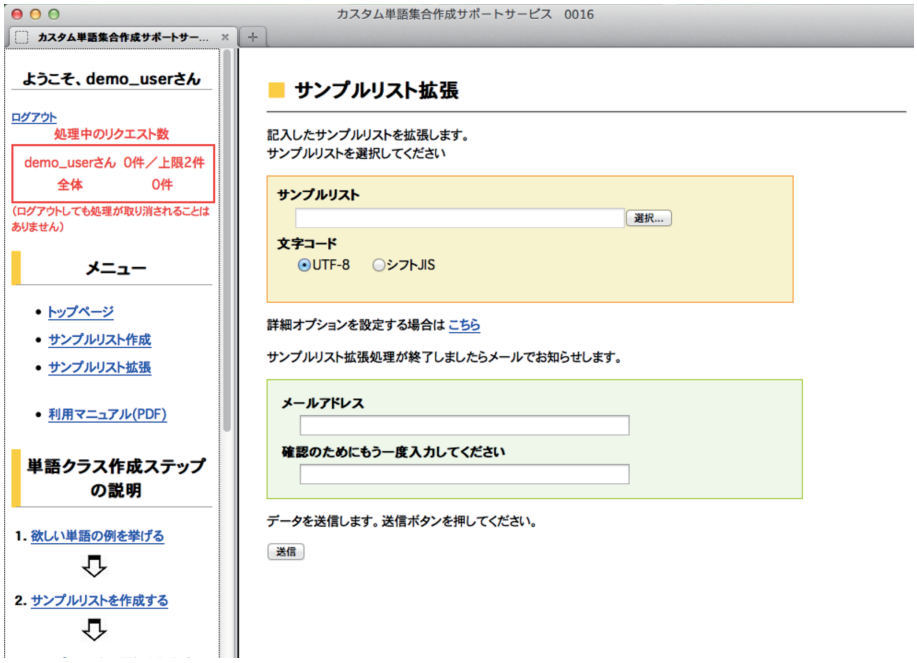

Fig.3 *User interface for Support Service for Customized Word Set Generation*

招　提　寺 (*toshodaiji*) / Toshodaiji Temple"...

- "釣り道具 (*tsuridogu*) / Fishing tackle" class
  - "釣り竿 (*tsurizao*) / fishing rod", "餌(*esa*) / bait", "ルアー (*rua*) / lure", "針(*hari*) / hook", "おもり (*omori*) / sinker", "テグス (*tegusu*) / fishing gut", "天　秤 (*tenbin*) / tenbin", "リ　ー　ル (*riru*) / reel", "竹　竿 (*takezao*) / bamboo rod", "玉 網 (*tamaami*) / landing net", "ルアーロッド (*ruaroddo*) / lure rod", "フライロッド (*furairoddo*) / fly rod", "釣　り　糸 (*tsuriito*) / fishing line", "タコテンヤ (*takotenya*) / octopus tenya", "ランディングネット (*randingunetto*) / landing net", "毛　針 (*kebari*) / feather hook", "アンカーロープ (*ankaropu*) / anchor rope", "人工餌 (*jinkoesa*) / synthetic bait", "さ　び　き (*sabiki*) / sabiki hook", "ジ　グ (*jigu*) / jig", "エギ (*egi*) / bait log", "テキサスリグ (*tekisasurigu*) / Texas rig", "ワーム (*wamu*) / worm", "餌木 (*egi*) / bait log", "カ　ッ　ト　テ　ー　ル (*kattoteru*) / cut tail worm", "仕掛 (*shikake*) / gimmick"...

The users of the service can interactively generate word classes on the browser-based interface shown in Fig. 3. To generate word classes of their own choice, they do not need any special expertise. All they need to do is to follow the instructions shown on the interface.

6.3 Semantic Relation Acquisition Service

"Semantic Relation Acquisition Service" is a Web based service that provides the users with word pairs that have a certain relation such as relations between "cause and effect", "trouble and preventive measure", "musician and song title", "location name and local specialty" and "hero and enemy". The service enables efficient semi-automatic generation of a large amount of word pairs having a specific relation using 6 hundred million Web documents based on a statistical method. Table 17 shows examples of word pairs that have "cause — effect" and "trouble — preventive measure" relations.

Users of the service can obtain semantic relations of their own choice just by inputting a few phrasal patterns that denote the relations. For example, if a user wants to get information about word pairs that have a causal relation, all he/she has to do is to input such phrases as "A causes B" and "A is the cause of B". The system then will automatically learn the patterns that may also have a causal relation such as "A triggers B" and "A generates B". Thus, the system keeps learning a great amount of similar patterns including those that are hard to think of for many people to provide the user with word pairs that have the semantic relation that the user wants to get by using all the possible similar patterns.

Since the system is designed to obtain a huge amount of semantic relations using various automatically learned patterns, it can found "unexpected but useful information" that are highly possibly overlooked by usual Web searches.

Like the case of "Support Service for Customized Word Set Generation", the users of the service can interactively generate word classes on the browser-based interface shown in Fig. 4. To acquire semantic relations of their own choice, they do not need any special expertise. All they need to do is to follow the instructions shown on the interface.

6.4 Parallel Search System for Similar Strings: Para-SimString

Written materials are one of the most familiar ways to deliver our messages to others. However, since they are written in natural languages, the same information is often conveyed by using different expressions, i.e. paraphrases, which may be one of the causes that hinder efficient management of documents and information. Unfortunately, technologies to recognize paraphrases at high speed among a large amount of documents have not been developed although automatic recognition of paraphrases has been actively studied. Para-SimString provides a means to retrieve paraphrases of certain expressions from a huge amount of documents in a fast and flexible way by narrowing down its targets to the ex-

Table 17 *Examples of "cause - effect" relation and "trouble - preventive measure" relation*

Cause - Effect	Trouble - Preventive Measure
連鎖球菌 (*rensakyukin*) - 化膿性関節炎 (*kanosei-kansetsuen*) streptococcus - septic arthritis	情報漏えい (*johoroei*) - 暗号化ソフトウェア (*angokasofutouea*) information leakage - encryption software
EB ウィルス (*EB uirusu*) - 伝染性単核球症 (*densen-seitankakukyuusho*) Epstein-Barr Virus - infectious mononucleosis	不正アクセス (*fuseiakusesu*) - ファイヤーウォール機能 (*faiyaworukino*) unauthorized access - firewall operations
ツボカビ (*tsubokabi*) - カエルツボカビ症 (*kaerutsubokabisho*) Chytridiomycetes - chytridiomycosis	床ずれ (*tokozure*) - エアマット (*eamatto*) bedsore - air mattress
断層 (*danso*) - 直下型地震 (*chokkagatajishin*) dislocation - epicentral earthquake	鳥害 (*torigai*) - 防鳥ネット (*bochonetto*) bird damage - bird net

pressions whose degree of similarity suffices a certain level as well as introducing parallel processing.

To be more precise, Para-SimString is a program to retrieve the lines that are superficially similar to the query string input by a user from a huge amount of document sets distributed on cluster computers in a high-speed and parallel way. For example, when a user input a query string "消費税の増税を閣議決定した (*shohizei no zozei wo kakugiketteishita*) / raise in the consumption tax was approved by the cabinet", Para-SimString retrieves lines as "消費税増税を閣議で決定 (*shohizeizozei wo kakugi de kettei* / consumption tax hike was approved by the cabinet)" and "消費税率増を内閣が決定した (*shohizeiritsuzo wo naikaku ga ketteishita* / the cabinet approved to increase the consumption tax rate)" from a large amount of documents if there are such lines there. In other words, it can comprehensively retrieve the strings that do not exactly match the query string but denote almost the same thing and are similar in their surface form.

Fig.4 *User interface for Semantic Relation Acquisition Service*

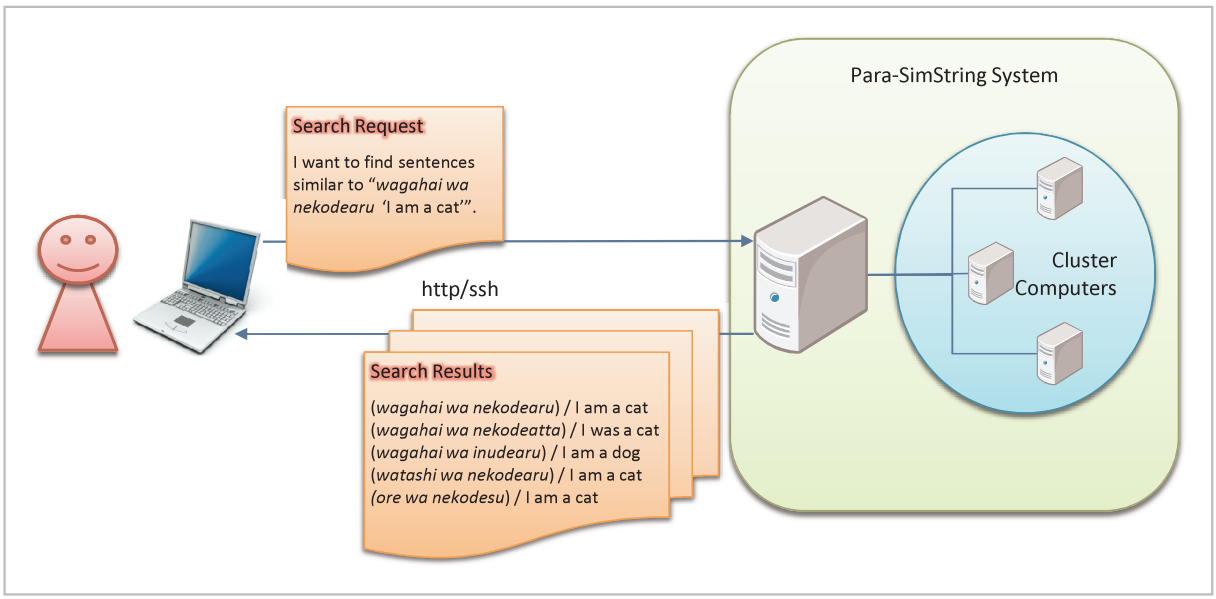

Fig.5 *Input/output flow and system configuration of Para-SimString*

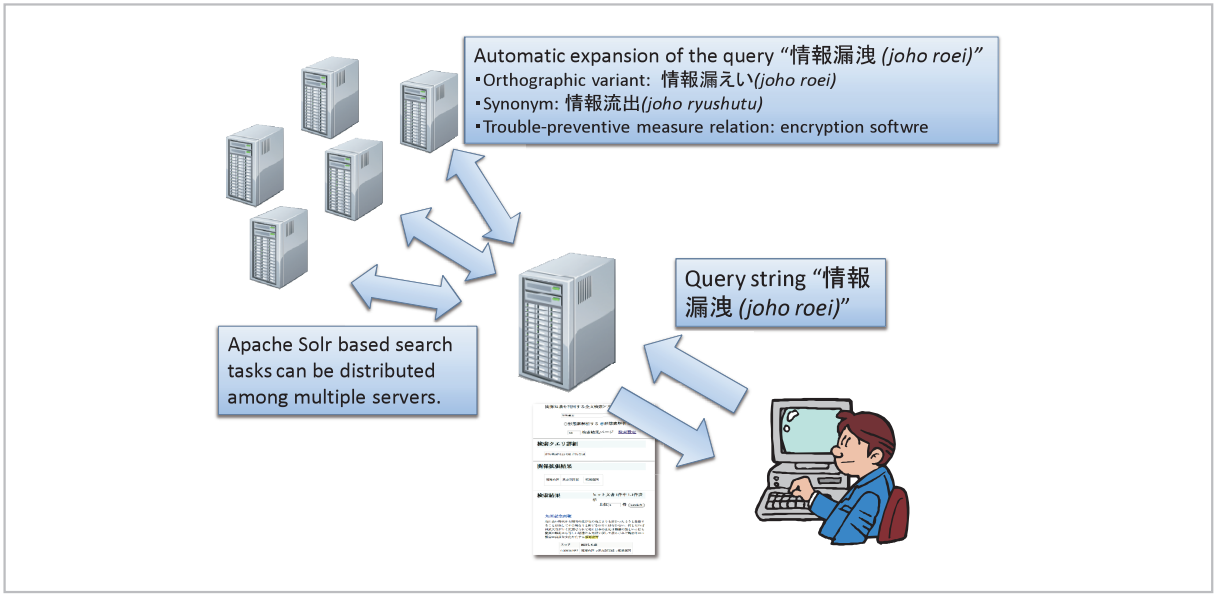

Fig.6 *System configuration of QE4Solr and an example of how QE4Solr expands queries*

What makes Para-SimString unique is its ability to perform parallel operations of indexing and retrieval. This is especially effective for handling enormous amounts of document sets and becomes an even more powerful advantage in parallel computing environments.

Para-SimString uses the open source software SimString[*6] for its core indexing and retrieval engine.

Figure 5 illustrates Para-SimString's input/output flow and its system configuration.

6.5 Query Expansion System for Solr: QE4Solr

To obtain desired information by searching documents accumulated in a commercial or academic organization often requires knowledge in the specific field where the organization is engaged. For example, when trying to search the documents held by an artificial in-

＊6 http://www.chokkan.org/software/simstring/index.html.ja

telligence-related department in a college, one may have to know that the terms "AAAI", "Association for the Advancement of Artificial Intelligence" and "アメリカ人工知能学会" all denote the same thing. QE4Solr is a query expansion system designed to run on the open source search platform Apache Solr. Knowledge bases can be used flexibly and easily on QE4Solr for automatic expansion of query strings. For example, incorporating a knowledge base containing information that explicitly denotes the specialty or singularity of a certain organization enables an intelligent search that matches the characteristics of that organization, or incorporating a knowledge base containing a large amount of orthographic variants, synonyms and semantic relations may prevent a search system from failing to find otherwise appropriate terms or provide us with unexpected but useful information.

Such knowledge bases can be easily constructed by utilizing Web-based services such as Support Service for Customized Word Set Generation and Semantic Relation Acquisition Service or other databases introduced in this paper.

QE4Solr's ability to perform parallel operations of indexing and retrieval enables an efficient search of large scale documents such as a Web archive.

Figure 6 illustrates how QE4Solr expands query strings and its system configuration.

7 Conclusion

In this paper, we have presented fundamental language resources constructed by Universal Communication Research Institute's Information Analysis Laboratory, including those that have not been published yet.

Fundamental language resources are building blocks for highly intelligent natural language information processing systems and important infrastructure that serves as a foundation to support the development of Japan's ICT technologies. However, construction of such resources requires a large amount of money for securing such resources as a large-scale parallel computing environment, many richly-experienced linguistic data annotators and researchers with expertise in information processing and many organizations have found it very difficult to raise the fund for securing such resources.

One of our missions is to contribute to the steady progress of Japan's ICT technologies including natural language information processing by continuously constructing and providing high-quality fundamental language resources including those that require a large amount of cost for construction, and we believe that our activities have made fundamental language resources greatly organized during the last few years.

Fundamental language resources must make further progress in their quality and quantity to contribute to the construction of natural language information processing systems that have almost human-level intelligence. In addition to the fundamental language resources that we have presented here, we have many more unreleased resources, and we believe that those resources will highly possibly lead to a technological breakthrough in the field of natural language information processing.

References

1 KAZAMA Jun'ichi, WANG Yiou, and KAWADA Takuya, " Fundamental Natural Language Processing Tools," Special issue of this NICT Journal, 5-4, 2012.

2 UCHIMOTO Kiyotaka, TORISAWA Kentaro, SUMITA Eiichiro, KASHIOKA Hideki, and NAKAMURA Satoshi, "Advanced Language Information Forum (ALAGIN)," Special issue of this NICT Journal, 8-1, 2012.

3 Jun'ichi Kazama, Stijn De Saeger, Kentaro Torisawa, and Masaki Murata, "Making a Large-scale synonym List using Stochastic Clustering on Dependency Structure," NPL 2009 (15th annual meeting of The

Association for Natural Language Processing), pp. 84–87, 2009. (in Japanese)
4 Kow Kuroda, Jun'ichi Kazama, Masaki Murata, and Kentaro Torisawa, "The accreditation criteria Japanese Orthographic Variant Pairs for Web Data," NPL 2010 (16th annual meeting of The Association for Natural Language Processing), pp. 990–993, 2010. (in Japanese)
5 Masahiro Kojima, Masaki Murata, Jun'ichi Kazama, Kow Kuroda, Atsushi Fujita, Eiji Aramaki, Masaaki Tsuchida, Yasuhiko Watanabe, and Kentaro Torisawa, "The Acquisition for Japanese Orthographic Variant Pairs in Short Edit Distance using Machine Learning and Various Features," NPL 2010 (16th annual meeting of The Association for Natural Language Processing), pp. 928–931, 2010. (in Japanese)
6 Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, and Masaki Murata, "Large scale relation acquisition using class dependent patterns," In ICDM '09: Proceedings of the 2009 edition of the IEEE International Conference on Data Mining series, pp. 764–769, 2009.
7 Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Takuya Kawada, Stijn De Saeger, Jun'ichi Kazama, and Yiou Wang, "Why question answering using sentiment analysis and word classes," In EMNLP, 2012.
8 Jun'ichi Kazama, Stijn De Saeger, Kow Kuroda, Masaki Murata, and Kentaro Torisawa, " A bayesian method for robust estimation of distributional similarities," In Proceedings of The 48th Annual Meeting of the Association for Computational Linguistics (ACL 2010), pp. 247–256, 2010.
9 Jun'ichi Kazama and Kentaro Torisawa, "Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations," In ACL-08: HLT: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 407–415, 2008.
10 Kow Kuroda, Jae-Ho Lee, Hajime Nozawa, Masaki Murata, and Kentaro Torisawa, "The Hand-crafted Cleaning for Hypernym Data of TORISHIKI-KAI," NPL 2009 (15th annual meeting of The Association for Natural Language Processing), pp. 928–931, 2009. (in Japanese)
11 Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki, "Enhancing the japanese wordnet," In The 7th Workshop on Asian Language Resources, 2009.
12 Kow Kuroda, Francis Bond, and Kentaro Torisawa, "Why wikipedia needs to make friends with wordnet," In Proceedings of The 5th International Conference of the Global WordNet Association (GWC-2010), 2010.
13 Patrick Pantel and Deepak Ravichandran, "Automatically labeling semantic classes," In HLT-NAACL '04: Proceedings of Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, pp. 321–328, 2004.
14 Masaaki Tsuchida, Stijn De Saeger, Kentaro Torisawa, Masaki Murata, Jun'ichi Kazama, Kow Kuroda, and Hayato Ohwada, "Analogy-based Relation Acquisition Using Distributionally Similar Words," IPSJ Journal, Vol. 52, 2011. (in Japanese)
15 Stijn De Saeger, Kentaro Torisawa, and Jun'ichi Kazama, "Looking for trouble," In Proceedings of The 22nd International Conference on Computational Linguistics, pp. 185–192, 2008.
16 Jun'ich Kazama, Stijn De Saeger, Kentaro Torisawa, Jun Gotoh, and István Varga, "Approach to the application of Question Answering System for Emergency Information," NPL 2012 (18th annual meeting of The Association for Natural Language Processing), pp. 903–906, 2012. (in Japanese)
17 Chikara Hashimoto, Kentaro Torisawa, Kow Kuroda, Masaki Murata, and Jun'ichi Kazama, "Large-scale verb entailment acquisition from the web," In Proceedings of EMNLP, pp. 1172–1181, 2009.
18 Chikara Hashimoto, Kentaro Torisawa, Kow Kuroda, Stijn De Saeger, Masaki Murata, and Jun'ichi Kazama, "Large-scale Verb Entailment Acquisition from the Web," IPSJ Journal, Vol. 52, No. 1, pp. 293–307, 2011. (in Japanese)
19 Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jun'ichi Kazama, and Sadao Kurohashi, "Extracting paraphrases from definition sentences on the web," In Proceedings of ACL/HLT, pp. 1087–1097, 2011.

20 Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jun'ich Kazama, and Sadao Kurohashi, "Paraphrasing Knowledge Acquisition of Definitions on the Web," NPL 2012 (17th annual meeting of The Association for Natural Language Processing), pp. 903–906, 2011. (in Japanese)
21 Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jonghoon Oh, and Jun'ich Kazama, "Another Semantic Polarity "Excitation/Inhibition" and Application to Knowledge Acquisition," NPL 2012 (18th annual meeting of The Association for Natural Language Processing), pp. 93–96, 2012. (in Japanese)
22 Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama, "Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web," In Proceedings of EMNLPCoNLL 2012: Conference on Empirical Methods in Natural Language Processing and Natural Language Learning (to appear), 2012.
23 Peter D. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002), pp. 417–424, 2002.
24 Hiroya Takamura, Takashi Inui, and Manabu Okumura, "Extracting semantic orientation of words using spin model," In Proceedings of the 43rd Annual Meeting of the ACL, pp. 133–140, 2005.
25 Julien Kloetzer, Stijn De Saeger, Kentaro Torisawa, Motoki Sano, Jun Goto, Chikara Hashimoto, and Jong Hoon Oh, "Supervised recognition of entailment between patterns," NPL 2012 (18th annual meeting of The Association for Natural Language Processing), pp. 431–434, 2012.
26 Takuya Kawada, Tetsuji Nakagawa, Ritsuko Morii, Hisashi Miyamori, Susumu Akamine, Kentaro Inui, Sadao Kurohashi, and Yutaka Kidawara, "The evaluation and classification for organize information and building a tagged corpus in Web text," 14th annual meetings of the Association for Natural Language Processing, pp. 524–527, 2008. (in Japanese)
27 Takuya Kawada, Tetsuji Nakagawa, Susumu Akamine, Ritsuko Morii, Kentaro Inui, and Sadao Kurohashi, "Tagging criteria of evaluation information," 2009. (in Japanese)
http://www2.nict.go.jp/univ-com/isp/x163/project1/eval_spec_20090901.pdf
28 Asuka Sumida and Kentaro Torisawa, "Hacking Wikipedia for hyponymy relation acquisition," In IJCNLP '08: Proceedings of the Third International Joint Conference on Natural Language Processing, pp. 883–888, Jan. 2008.
29 Jong-Hoon Oh, Kiyotaka Uchimoto, and Kentaro Torisawa, "Bilingual co-training for monolingual hyponymy-relation acquisition," In ACL-09: IJCNLP: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp. 432–440, 2009.
30 Stijn De Saeger, Jun'ichi Kazama, Kentaro Torisawa, Masaki Murata, Ichiro Yamada, and Kow Kuroda, "A web service for automatic word class acquisition," In Proceedings of the 3rd International Universal Communication Symposium, pp. 132–138. ACM, 2009.

(Accepted June 14, 2012)

***HASHIMOTO Chikara**, Ph.D.*

Senior Researcher, Information Analysis Laboratory, Universal Communication Research Institute

Natural Language Processing

***OH Jong-Hoon**, Ph.D.*

Researcher, Information Analysis Laboratory, Universal Communication Research Institute

Natural Langauge Processing

***SANO Motoki**, Ph.D.*

Researcher, Information Analysis Laboratory, Universal Communication Research Institute

Linguistics

***KAWADA Takuya**, Ph.D.*

Researcher, Information Analysis Laboratory, Universal Communication Research Institute

Linguistics