École Nationale Supérieure des Mines de Paris
Corps Techniques de l'État

Overview of Research in Written Language Processing in Japan

Jean-Philippe Vert
Ingénieur des Mines

July, 1998

Contents

1 Introduction **7**

2 Morpho-syntactic analysis **9**
2.1 Segmentation and morphological analysis 9
2.1.1 JUMAN at Kyoto University 9
2.1.2 A statistical approach at the NAIST 10
2.1.3 Decision trees without dictionaries in ATR 10
2.2 Syntactic parsing 11
2.2.1 KNP : a parser for long sentences, developed at Kyoto University 11
2.2.2 An English parser based on decision trees at ATR . . 12
2.2.3 HPSG formalism in Tokyo University 12
2.2.4 POWER, an object-oriented parser in Niigata University 13
2.2.5 A probabilistic syntactic parser in Tokodai 13
2.3 Combined analysis 14
2.3.1 A morpho-syntactic parser based on LR algorithm in Tokodai 14

3 Semantic analysis and representation **15**
3.1 Structured organizations of concepts 15
3.1.1 A concept dictionary at EDR 15
3.1.2 Co-occurrence graphs at ETL 16
3.1.3 Hierarchical clustering of words at ATR 16
3.2 Non-organized semantic representation 17
3.2.1 A vectorial approach for information extraction at NTT 17
3.2.2 Semantic vectors at Shinshu University 17
3.2.3 A semantic representation using neural networks at ETL 18
3.3 Mixed approaches 18
3.3.1 Automatic classification of documents at Tokushima University 18
3.3.2 Words similarity measurement at NTT CS Laboratory 19

4 Applications **21**
4.1 Information retrieval and extraction 21
4.1.1 Request expansion at Tokodai 22
4.1.2 Geographical knowledge organization at NTT 22
4.1.3 5W1H classification at NEC 23
4.1.4 Information extraction with template matching at Kyutech 23
4.1.5 Use of a concept base at NTT 24
4.2 Document classification 24
4.2.1 Kanjis retrieval at Kyoto University 24
4.2.2 Automatic document classification at Tokushima University 25
4.3 Multimedia alignment at Kyoto University 25
4.4 Characters disambiguation 25
4.4.1 N-gram models at Kyoto University 26
4.4.2 A mixture of n-gram models at Tohoku University . . 26
4.5 Automatic Summarization 26
4.5.1 The system TESS at Kyutech 26
4.6 Machine translation (MT) 27
4.6.1 NEC's PIVOT 27
4.6.2 NTT's ALT-JE 28
4.6.3 ATR's MATRIX 28
4.7 Human/machine interface 28
4.7.1 A human/machine dialogue system at NTT 29
4.7.2 The competition DIALEAGUE organized by ETL . . 29
4.7.3 IRENA system at Kyutech 29

5 Linguistic resources **31**
5.1 EDR's productions 31
5.2 NTT's productions 32
5.3 Kyoto University corpus project 32
5.4 ATR's resources 33
5.5 ETL's GDA project 33

6 Laboratories **35**
6.1 Advanced Telecommunications Research Institute International (ATR) 35
6.2 Electrotechnical Laboratory (ETL) 35
6.3 Japan Electronic Dictionary Research Institute, Ltd (EDR) . 37
6.4 Kyoto University 37
6.5 Kyushu Institute of Technology (Kyutech) 37
6.6 Nara Institute of Science and Technology (NAIST) 38
6.7 NEC 38
6.8 Niigata University 38
6.9 NTT Basic Research Laboratories (BRL) 39

6.10 NTT Communication Science Laboratories 39
6.11 NTT Human Interface Laboratories 39
6.12 NTT Software Laboratories 39
6.13 Shinshu University 40
6.14 Tohoku University 40
6.15 Tokushima University 40
6.16 Tokyo Institute of Technology 41
6.17 Tokyo University 41

Chapter 1

Introduction

Natural Language Processing (NLP) took advantage of the apparition of powerful computers together with the development of new tools and methods, involving linguistics, mathematics and computer science to make important progress recently. Even though many challenges remain to be taken up, like automatic translation or virtual operators communicating with humans in natural language, it appears that the advances in this field are of huge economical interest. As a result, large companies are investing a lot of money and trying to create research networks in this field, in order to obtain results as quickly as possible. Institutions like the DARPA in the United States, the MITI in Japan or the European Commission are also funding many large-scale programs to foster research.

This report gives an overview of the research carried out in Japan in NLP, and limits itself to written language processing. It stresses the research topics of interest in 1998 and presents the laboratories involved. It was achieved after the study trip made by a student of the Corps National des Mines (France) who had the opportunity to meet many Japanese researchers and to visit their laboratories.

This work was made possible by a grant from the French Ministry of Economy, Finance and Industry, and the extreme kindness and open-mindedness of all the Japanese laboratories that were solicited.The author thanks them all.

It begins with a presentation of the tools developed to analyse written language, including morphological and syntactic analysis. It then focuses on the semantical representation of language and the main applications currently under development. A section is also devoted to the linguistic resources available, followed by a quick presentation of the laboratories who contributed to this report by inviting its author to visit them.

The goal of this report is to give an overview of the research. Therefore the author chose to write a report for a non-specialist, often at the expense of the scientific precision that a deeper study of any research topic would

need. The interested reader is therefore invited to visit the web sites of the laboratories indicated in the last part, which often contain a bibliography, or to contact the author of this report at the following e-mail:

jean-philippe.vert@mines.org

Chapter 2

Morpho-syntactic analysis

Before manipulating texts written in natural language, it is almost always necessary to carry out a first analysis of the sequence of characters in order to obtain the following kind of information:

- what are the basic linguistic entities (segmentation)?
- what is the morphological nature of these entities (morphological analysis)
- what are the grammatical structure of the sentence and the syntactic characteristics of each word (syntactic analysis)

2.1 Segmentation and morphological analysis

The segmentation task is particularly important in Japanese, as no space separates the words. This task and the morphological analysis of the text are often grouped together into one single operation, by using a lexicon that contains a list of all possible sequences of characters together with their morphological description. Given a text, i.e. a sequence of characters, that lexicon is used to obtain a list of possible segmentations and morphological analysis. Constraints are then applied to eliminate possibilities in this list, or to classify the candidates according to their probabilities.

2.1.1 JUMAN at Kyoto University

JUMAN is a morphological parser for Japanese that was developed in Kyoto University by the team of Prof. Nagao. When a sentence is presented to JUMAN, it segments it into morphems and indicates the morphological class of every morphem (e.g. name, verb...). Two dictionaries are used to obtain this result:

- a *grammar dictionary*, that contains the list of morphem classes and subclasses, a conjugation dictionary, and a set of connection rules between morphems;

- a *morphem dictionary*, that contains the list of morphems used in Japanese together with several informations (class and sub-class, pronunciation, meaning...)

One advantage of JUMAN is its modularity and adaptability. Indeed, in order to be able to deal with a large number of different morphological formalisms or grammars, it was conceived as a kernel that uses the dictionaries of any user. Even though it is given with a complete configuration (including a 120,000 words dictionary and a list of 14 morphological classes), it can be easily adapted to any personal formalism and dictionary. In Kyoto University, it is currently used with the 230,000 entries EDR dictionary and deals with 3,000 classes of morphems. In the case where several parsing candidates exist for one sentence, JUMAN prefers the one that contains the smallest number of unknown words, morphems and independent words.

JUMAN is used in many laboratories as morphological parser, in Japan as well as abroad. It is free and can be downloaded from the Internet from Dr. Kurohashi's laboratory in Kyoto University, or from Prof. Matsumoto's laboratory in NAIST.

2.1.2 A statistical approach at the NAIST

Instead of using a fixed grammar for morphological parsing, Prof. Matsumoto's laboratory, in NAIST, has developed a statistical approach. The parser first uses a dictionary to obtain a list of all possible candidates for parsing. A cost is then associated to every bigram, depending on the frequencies measured on a corpus. The lowest cost analysed is finally considered to be the most probable one. The formalism used by the parser consists of 14 different morphological classes, divided into sub-categories (e.g. common names, country names, etc...). The bigram costs have first been estimated by hand, but a computation using EDR corpus was in development in July 1998.

2.1.3 Decision trees without dictionaries in ATR

At the Advanced Telecommunications Research Institute (ATR), researchers have pointed out the proliferation of tagging systems and associated dictionaries for the morphological analysis of Japanese. They have therefore developed a *robust* parser based on decision trees and that does not require any dictionary, in order to overcome the following frequent problems:

- words that appear in texts and that are not in the dictionaries;

- words that are in the dictionaries, but not with the correct part-of-speech.

This method consists in creating a large number of questions, whose answers should contain information about the words boundaries and POS. For example, the strings "ing" at the end of a word in English might indicate that the word is a verb. In the same way, many structures in Japanese can give information about the morphology of texts. Such observations result in a series of questions that can combine several points of interest (e.g. "the word ends with *hiragana* characters and the following character is a *kanji*").

The resulting series of questions is then used to create decision trees with classical methods, thanks to a training step done with an already annotated corpus. The tag sets used in the experiments are composed of 209 different tags, grouped in 18 classes (common noun, proper noun etc ...).

Once trained, each final leaf of the decision tree contains a probability distribution for the morphological structure (segmentation and tagging) of input sentences.

2.2 Syntactic parsing

The syntactic parsing of a sentence consists in finding the correct syntactic structure of that sentence in a given formalism. Formalisms are called *grammars*, and contain the structural constraints of language. Whereas morphological analysis of Japanese can be considered as very specific and different from European languages, the problem of syntactic parsing is very similar for Japanese and other languages.

2.2.1 KNP : a parser for long sentences, developed at Kyoto University

The *Kurohashi-Nagao Parser* (KNP) is an algorithm specially developed for parsing long sentence, which is classically a complex problem if relations exist between words far away from each other. The basic assumption of the algorithm is that a long sentence usually contains *conjunctive structures* that link noun phrases or clauses with each others. For example, in the sentence "Paul liked the book and enjoyed the movie" the word "and" links two parallel clauses. This kind of structure is very common in Japanese and creates many ambiguities for syntactic parsers.

In order to discover these conjunctive structures, members of Nagao Laboratory in Kyoto University have proposed the KNP algorithm that computes the *similarity between two sequences of words* on the left and on the right of a conjunction, and that selects those series of words that can be considered as similar enough to belong to a conjunctive structure, using technics of dynamic programming.

The similarity computation between blocks of words is the key of the algorithm and is based on comparisons between morphological categories, Japanese characters used, and semantic classes of the words in different blocks obtained from a thesaurus.

2.2.2 An English parser based on decision trees at ATR

Using the "ATR English Grammar", researchers from ATR have developed a syntactic parser that is based on a 3000-tags and 1100-rules grammar. The syntactic structure of a sentence is written as a tree, every non-terminal node of which contains the identification of the rule that created it, and every terminal node of which contains the POS of the corresponding word. A tagged corpus using this grammar has been created in collaboration with Lancaster University (the "ATR/Lancaster Treebank of General English") and has been used for the training of decision trees that automatically build the tree associated with a sentence. The questions used in the decision tree concern the attributes of the terminal and non-terminal tree nodes, as well as other characteristics of words or sentence (e.g. the size of the sentence).

An analysis is the result of a succession of partial analyses, that represent successive states in the process of tree determination. The jump from one state to another happens when a new node is tagged, or when a new node is set to be terminal : these decisions are taken thanks to decision trees.

2.2.3 HPSG formalism in Tokyo University

Tsujii Laboratory in Tokyo University is working on the development of a framework and an environment based on Head-Driven Phrase Structure (HPSG) formalism, in order to enable applications of NLP like knowledge acquisition, machine translation or information extraction.

HPSG formalism is almost not used for practical applications because it is less efficient that other approaches, using for instance context-free grammars. However, it has several advantages like minimizing the size of the grammar, which only contains 6 rules for the Japanese grammar used in the laboratory of Professor Tsujii, and maximizing the role of lexical information.

The parser uses a two-step structural analysis:

- first analysis enumerates possible structures, using CFG rules compiled from HPSG lexical rules;
- then the parser solves the constraints that are not contained in the CFG grammar.

Several points were under development in 1998 in order to foster the use of HPSG formalism:

- creation of finite-state automata;

- creation of parallel algorithms to increase the speed;
- information extraction tools;
- automatic creation of a thesaurus.

The research on HPSG involves several collaborators, especially in Japan (e.g. EDR), in the USA (e.g. Stanford University) or in Europe (e.g. DFKE in Germany).

2.2.4 POWER, an object-oriented parser in Niigata University

Miyazaki Laboratory in Niigata University has been working for several years on the development of a new type of syntactic parser for Japanese and English : POWER. It is an object-oriented parser written in PROLOG.

The basic units that the program is working with are objects, that are made of a word together with a class (e.g. noun) and a category (e.g. subject). Depending on their attributes, objects send messages to other objects, which carry out different actions and answer. Verbs will, for instance, send messages to subjects and wait for answers.

POWER deals with Japanese sentences that have already been morphologically parsed, using JUMAN for instance. Structure trees are sorted by costs that are linked with their probability of being the correct structure of the sentence. One advantage of POWER is its flexibility with respect to the formalism used and the actions that the objects have to carry out.

2.2.5 A probabilistic syntactic parser in Tokodai

Tanaka Laboratory in the Tokyo Institute of Technology (Tokodai) has generalized the method called GLR (Generalized Left-Right) to include a probability estimation for every candidate obtained from this parsing method. In case of ambiguities between several candidates, the one with the highest probability can then be preferred, and an information concerning the reliability of the result can also be obtained from this method. Besides, the speed of analysis is increased by eliminating candidates with very low probability before they are completely estimated.

This method is more precise than Probabilistic Context-Free Grammar (PCFG) which consists in estimating the probability of each grammar rule. In order to get a mildly context-sensitive model, probabilities are estimated for actions of the LR table used by the automat to parse sentences and derived from the CFG rules. The probability of a derivation is defined as the product of the probabilities of the actions that led to that derivation. Probabilities are estimated by counting frequencies of actions to parse a corpus of correctly parsed sentences, in this case 10,000 sentences from the ATR corpus.

2.3 Combined analysis

Whereas morphological and syntactic analysis are often separated into different stages, the second one using the results of the first one, some researchers try to combine these two steps into one global parsing, justifying this approach by the fact that informations of each these two steps are often useful for the other one, especially in the case of ambiguities.

2.3.1 A morpho-syntactic parser based on LR algorithm in Tokodai

The research team of Professor Tanaka in the Tokyo Institute of Technology proposed a method to combine morphological and syntactic analysis in one stage and to keep distinct morphological and syntactic rules. That point is very important, as much research has already been done in these two fields that led to efficient models.

Classical morphological analysis uses a dictionary that specifies the morphological category *mcat* of every word or series of characters, and a connectivity matrix in order to enable or not a sequence of *mcat*. Using this approach alone usually leads to many ambiguities.

To overcome that problem Professor Tanaka's team combines this classical approach with syntactic LR parsing that derives a LR matrix used by an automat from CFG rules. This is done by adding to the CFG rules (that concern syntactic categories *cat*) a set of rules to link *mcat* with *cat* thanks to the dictionary that gives for each word its *cat* and *mcat*. One *cat* is usually associated to several *mcat*. The grammar is automatically augmented and is then considered as a CFG from which a LR table can be derived, based on a set of *mcat* and *cat*.

The connectivity matrix is then used to automatically delete illegal reduction actions, in order to obtain a modified LR table that includes morphological constraints. A modified LR parsing algorithm can finally be applied to sentences in order to build a tree that sums up the morphological and syntactic structures.

Chapter 3

Semantic analysis and representation

The semantic analysis of a word, a sentence or a text, consists in determining its meaning. The notion of meaning is nevertheless ambiguous and can only be precisely defined with respect to a given formalism that is called *semantic representation*. The current chapter aims at presenting the formalisms developed and used by Japanese laboratories.

Two main families of semantic representations can be observed:

- structured organizations of semantic units, called concepts, usually with the help of graphs;
- non-sorted series of concepts that define a high-dimension vectorial space whose vectors are meanings. This representation enables simple tools of Euclidean geometry to be used, e.g. innerproduct or projections.

3.1 Structured organizations of concepts

3.1.1 A concept dictionary at EDR

The electronic dictionary created by the Japan Electronic Dictionary Research Institute, Ltd. (EDR) is made of several dictionaries, one of which is a concept dictionary. This dictionary describes the relationships between 400,000 concepts that have been introduced as word meanings from the word dictionary. Thus, each concept is one particular word meaning, and is represented by a sequence of 5 alphanumeric characters. Each word of the Japanese and English dictionaries is therefore connected to one or several concepts.

The concept dictionary contains:

- a concept list, with their identification number, one illustration of their meaning through an example sentence, and an explanation of their meaning;

- a concept classification that describes the "super/sub" binary relationships between concepts;

- concepts descriptions, that describe binary semantic relationships between concepts, e.g. agent-object or object-action.

Using these dictionaries for NLP tasks usually involves tools to navigate inside graphs, in order to understand the meaning of sentences written in natural language. Semantic disambiguation can for instance be obtained by searching the positions of ambiguous words in the concept graph, and by comparing these positions.

3.1.2 Co-occurrence graphs at ETL

Doctor Tanaka-Ishii from the Electrotechnical Laboratory (ETL) in Tsukuba has worked on building words co-occurrence graphs in order to get a representation of concepts. That approach was first aimed at comparing a non-aligned multi-lingual corpus. Indeed, with the help of a bilingual dictionary and co-occurrence graphs it is possible to find a mapping from one graph to the other that can be used for disambiguation of words in the framework of machine translation.

The graph is thus automatically built and is supposed to represent relationships between concepts, no matter what the language is, just like the EDR dictionary does.

3.1.3 Hierarchical clustering of words at ATR

Using a textual database made of articles from the *Wall Street Journal* researchers from the Advanced Telecommunications Research Institute (ATR) near Kyoto have obtained a hierarchical classification of the 70,000 words used the most often. The result of this classification is a binary tree where the 70,000 terminal leaves represent the 70,000 words, and where each node represent a class of words that contains the words of the children nodes.

The tree was built automatically, starting from 70,000 isolated leaves and clustering iteratively the classes used in similar contexts. Further, every node of the tree and thus every concept can be coded by a series of bits.

3.2 Non-organized semantic representation

3.2.1 A vectorial approach for information extraction at NTT

NTT's Human Interface Laboratory at Yokasuka, between Tokyo and Yokohama, was working on a project of topic extraction from broadcasted English news. It is based on a list of 70,000 words that describe topics, this list being therefore a concept base for a vector space representation of concepts.

Using information records a distance matrix was created between words of this list and any English word, by counting co-occurrences between the words of the list appearing in the headlines and all the words being used inside the news. The resulting matrix could then be used to create mutual information or χ^2 models, in order to finally define a score between each word of the list and each English word. The score of a word from the list with respect to a news article can thereafter be computed as the normalized sum of the scores of this word with respect to the words composing the news.

This model can be used to extract the words of the list that have the highest scores with respect to a given news report, in order to automatically obtain its topic. Conversely it is possible to retrieve the news article with respect to which a given word has the highest scores.

This method implicitly makes use of a 70,000-dimensional vector space representation of concepts, in which every English word is represented through its scores. Geometric tools naturally appear, as the use of average to define the scores of a series of words.

3.2.2 Semantic vectors at Shinshu University

Nakano, Okamoto and Maruyama laboratories at Shinshu University have worked on 3D representation for document retrieval on the Internet. They used semantic vectors defined as normalized vectors whose coordinates are the weights of the keywords used to define the basis of the vector space. These weights are computed from absolute and relative frequencies of the keywords in the document considered, compared with other documents of the database.

Those semantic vectors are used to represent:

- the semantic content of a document;
- a user's viewpoint.

They are also used to easily define the similarity between documents as the innerproduct of their semantic vectors. Simple technics (innerproducts, projections ...) can therefore have a meaning on a semantic point of view, when they are applied to semantic vectors.

3.2.3 A semantic representation using neural networks at ETL

Doctor Takahashi from ETL laboratories proposed a method to represent semantic contents of Japanese words or sentences with real vectors, thanks to neural networks.

These vectors, called *semantic representation vectors (SRV)* have all a fixed size and are obtained thanks to recursive auto-associative memories (RAAM) neural networks trained on a corpus, in order to assign similar SRV to similar words or sentences. Moreover the transformation from a sentence to a SRV is reversible so different sentences have different SRVs.

The semantic representation is thus obtained in a vector space whose base vectors are not defined by a list of concepts but automatically adjusted by neural networks.

3.3 Mixed approaches

3.3.1 Automatic classification of documents at Tokushima University

Aoe laboratory at Tokushima University, on Shikoku's island, has developed a system for automatic classification of textual documents in Japanese located in a folder, that will soon be commercialized. The classification of carried out with respect to the content of the documents to be classified, and is done in a two-steps process:

- retrieval of *keywords* in the documents;
- classification of documents using a hierarchy of concepts.

The keyword retrieval in a document is obtained by counting absolute and relative frequencies of a series of 40,000 character bigrams, to extract the ones that offer the best characterization for the document considered. That step can thus be considered as typical of vector space representations.

The second step however uses a semantic hierarchy on keywords in order to obtain a hierarchical classification of the set of documents itself. This step is therefore typical of structured concept representation.

The combination of these two approaches in order to classify a textual database with respect to the semantical content of the documents has the advantage of making use of computationally efficient tools through the vector representation, and integrating much semantic information with the pre-existing hierarchy of keywords.

3.3.2 Words similarity measurement at NTT CS Laboratory

NTT Communication Science (CS) Laboratory, located near the ancient imperial capital Nara, has created a hierarchy of about 3,000 semantic concepts linked to each others through "has-a" and "is-a" relationships. Even though it might appear much smaller than the 400,000-concepts EDR dictionary, it aims at being more robust than its elderly sister.

This concept base, called "knowledge base", is used as a base for a 3,000 dimension vector space in which all Japanese words are represented as vectors. The set of all Japanese words has first been reduced to a 40,000 concept-words set after standardization by NTT's thesaurus. The result is therefore a 3,000 concepts $\times$ 40,000 concept-words matrix, in which concept-words coordinates are normalized.

This construction enables the computation of two words similarity with respect to a particular viewpoint. With respect to the viewpoint "animal", for instance, the word "horse" is closer to the word "rabbit" than to the word "car", but the result is in the reverse order with respect to the viewpoint "transportation". To take care of the viewpoint in the measure of similarity, vectorial operations are used, e.g. projections to project a vector on a viewpoint and innerproduct to measure similarities.

Although NTT's approach looks like a purely vector space representation, one should notice that the base of the vector space is composed on concepts that are organized in a semantic graph, which open new fields of investigation for deeper analysis.

Chapter 4

Applications

Morpho-syntactic and semantic analysis of texts, like the ones presented in the previous chapters, are usually performed in order to prepare the text for an precise application. The goal of the current chapter is focus on these applications of NLP.

Just like in the USA or in Europe, the main efforts in Japan are focused on applications that will enable the computers to have a high added value compared with human in the field of textual document manipulation and processing. Among these applications with huge potential economical rewards we can quote the following:

- information retrieval and extraction in a large textual database;
- automatic classification of documents;
- alignment of several information sources (e.g. newspapers, television, radio...);
- disambiguation of speech or handwriting recognition;
- automatic summarization of texts;
- machine translation (MT);
- human/machine interfaces.

4.1 Information retrieval and extraction

Information retrieval (IR) in large textual databases is a very active field where competition is harsh and progress rapid. The discipline has known a breakthrough with the creation of the Text Retrieval Conference (TREC) held every year since 1991 in the USA with grants from the Defense Advanced Research Projects Agency (DARPA) and the National Institute of Standards and Technology (NIST). In 1997, 51 groups from 12 countries participated

in the seventh TREC conference. Whereas only one Japanese group (NEC) presented a result for the contest, several researchers from different research institutes in fact participated to this team's efforts. A similar competition is to be organized in Japan from 1999, but the textual database will consist of Japanese texts.

4.1.1 Request expansion at Tokodai

Tokunage laboratory at the Tokyo Institute of Technology (Tokodai) was working to develop a method of query expansion. When a user is looking for texts in a large database, he has to formulate his desire through a request, which is a series of words. A request expansion consists in completing the user's request with other words before sending it to usual search engines.

The method proposed by the laboratory uses the database called *Wordnet* that was developed from 1985 in Princeton University under the direction of Professor George Miller. This lexical knowledge source has been widely used for NLP applications, but has never produced wonderful results for information retrieval. The request expansion proposed by Professor Tokunaga's team consists in searching for new words similar to the ones entered by the user in his request thanks to three information sources‘:

- a thesaurus, automatically built from words co-occurrences;
- a thesaurus, automatically built from "predicate-argument" type data observed on a parsed corpus;
- the Wordnet database.

An experiment to test this expansion has been done in combination with Cornell University's *SMART* search engine. It showed that the combination of these three sources increased the performances of the search engines, and proved that the sources were complementary.

4.1.2 Geographical knowledge organization at NTT

NTT Software Laboratory, located in Musashino in the eastern suburb of Tokyo was installing on the Internet a knowledge search engine depending on the geographical position of the user. This tool could be reached at the URL http://www.kokono.net and was aimed at providing a guide of institutions, hotels, restaurants or stores located near the user, using information available on Internet. Therefore, as a first step, the software has to localize the user, using information from the telecommunication company, and, as a second step, pertinent information has to be retrieved from the Internet.

The method chosen for the second step was to organize geographically the information on Internet, either using databases like yellow pages in which

geographic informations are easy to extract, or by extracting such information directly from documents available on the Internet (e.g. ZIP codes or addresses).

This information extraction was therefore limited to geographic information, in order to organize a very large database according to an original criterion. The Internet site kokono.net was experimental in July 1998, but was about to become commercial.

4.1.3 5W1H classification at NEC

The*Pattern Analysis and Human Language Technology* group of NEC Corporation, a giant of computer manufacturing and communication, has developed a navigation engine for textual databases with respect to 5W1H requests (who, when, where, what, why, how). During the stage of information organization in the database the program extracts a 6-dimensional vector for every sentence, corresponding to the 6 elementary questions. This information extraction uses NLP and pattern matching techniques, in order to identify pertinent informations.

The navigation stage then consists of asking the user to fill in one or several fields of the 6-dimensional question vector, and searching for documents that correspond to the query. The applications tested in July 1998, concerned economic news for which 5W1H formalism is particularly well adapted.

4.1.4 Information extraction with template matching at Kyutech

Nomura laboratory at the Kyushu Institute of Technology has developed a software intended for information extraction from written news. A demonstration was available on the Internet in July 1998. It worked with a database of 2,000 news articles that dealt with the commercialization of new goods, from which several pieces of information could be extracted, e.g. type of goods, name, manufacturer, price etc ...

A morphological parsing of all texts was first carried out with the parser JUMAN developed by Kyoto University (see 2.1.1). The surface information provided by this analysis was then processed by the information search engine through template matching. In other words, any type of information was characterized by a series of templates concerning the linguistic environments of the information concerned - e.g., the Japanese particle that is used before a date - and the information itself. The search engine finds candidates for each piece of information through a classical template matching, by applying every template to every word of a given news article. This leads to the production of candidates for every sentence, and the final candidates selected by the search engine will be chosen according to the number of templates that selected each candidate and the number of sentences that extracted

them.

A set of almost 4,000 templates was first created in order to characterize the information to be extracted. A reduction procedure using a corpus then reduced that number to about 1,400. The precision of extraction appears to be over 90 percent correct extraction in the public demonstration.

4.1.5 Use of a concept base at NTT

NTT Communication Science (CS) Laboratory had an information retrieval project under development in collaboration with Stanford University and the Stanford Japan Center. This retrieval system is based on the concept base (see 3.3.2) that enables the measurement of similarity between words and the clustering of texts according to their content.

The concept base that gives information on the meaning of words contains about 20,000 concepts. Any set of words, e.g. a request, can be represented by a vector in the 20,000-dimensional concept vector space. Similarity between two words or sets of words is defined as the innerproduct of their representative vectors, and the vector of a set of words is defined as the average of the vectors of the words. Conversely the word characteristics for a vector of the concept space are the words whose vectors are close to the vector considered with respect to Euclidean norm.

These definitions can be used to expand a request, by adding those words that are characteristic of the vector representing the request. Moreover, articles with similar vectors can be clustered and characteristic words can be extracted from each class. In particular it is possible to retrieve texts that don't explicitly contain the words of the query but are semantically close to it.

While only a Japanese version was available during the visit of NTT's laboratory an English version has been developed at Stanford University.

4.2 Document classification

The goal of automatic document classification is to perform a clustering of texts in a database depending on their content.

4.2.1 Kanjis retrieval at Kyoto University

Doctor Kurohashi's laboratory at Kyoto University proposed a method for classifying Japanese documents without performing any morphological parsing of these documents but just by observing the *kanjis*[1].

[1] The Japanese characters called *kanjis* often have an intrinsic meaning even when then combined with other characters to form a word. Observing the kanjis used in a text therefore gives semantic information about the text

The model was trained using a database of texts already classified according to their topic (philosophy, architecture etc ...) to extract the kanjis characteristic for each topic using a χ^2 method. The kanjis found to be characteristic can thereafter be used to classify new texts depending on the kanjis observed in it.

4.2.2 Automatic document classification at Tokushima University

Aoe Laboratory at Tokoshima University has developed a software for automatic classification of textual documents in all folders. The software was still in development in July 1998, but was about to be released commercially. It can classify documents in a personal computer or on the Internet, can be used for documents in Japanese or in English, and returns a sentence that explains the content of each document class after classification.

The system uses keyword retrieval in texts (see 3.3.1) and the associated concept hierarchy to classify texts according to their content.

4.3 Multimedia alignment at Kyoto University

One multimedia application of NLP consists of automatically aligning emissions of different media depending on their content. The goal is for instance to detect when the same information in processed on TV, on the radio and in the newspaper.

Doctor Kurohashi's laboratory at Kyoto University has developed a system that automatically aligns news from TV broadcasts and newspapers. The system detects the words that appear in the different media and multiplies their frequencies by a weight that depends on their position (e.g. in the title, the headlines, the first paragraph of the text etc ...). By this method, an alignment is obtained between newspaper articles and TV programs having many affinities.

4.4 Characters disambiguation

Character disambiguation is widely used in automatic character recognition (O.C.R.) softwares as it consists of choosing one candidate among a list of possible characters that could correspond to a written sign. This task is usually carried out using a simplified model language that is built to give the probability of every character after a series of characters, often estimating such probabilities with n-gram models where n is usually equal to 2 or 3.

4.4.1 N-gram models at Kyoto University

In order to estimate probabilities with a n-gram model from a corpus, for any n, one needs to count the occurrences of every n-gram in the corpus, which can be computationally inefficient. Doctor Kurohashi's laboratory at Kyoto University uses a simple method to get these estimates quickly for any n.

The method consists of assigning one pointer to every character of the corpus used to train the model. The set of pointers is then sorted with respect to a lexicographic order on the sequences of characters that follow every pointer's character. Once sorted, the number of occurrences of any n-grams can easily be computed by counting the number of pointers that point on a character followed by the full n-gram, as these pointers are sorted.

4.4.2 A mixture of n-gram models at Tohoku University

Aso laboratory at Tohoku University in Sendai is specialized in intelligent digitalization of paper documents. It has in particular developed a character recognition system including a disambiguation engine.

This engine uses n-gram models as approximations of language models, but is original in so far as it mixes the models with $n = 0, 1, 2, 3$. As a result the final model is a linear combination of these different models where the weights for each models are adjusted with respect to the size of the training corpus. This enables the model to get approximations involving 3-grams even when the corpus is too small to obtain a consistent 3-gram model, as the weights are estimated in order to maximize the consistency of the final model.

4.5 Automatic Summarization

Automatic summarization is a typical NLP application that could enable computers to significantly improve human work's productivity, by replacing the reading of large quantities of documents by the reading of their summarization.

4.5.1 The system TESS at Kyutech

Nomura Laboratory at the Kyushu Institute of Technology has developed a software of automatic translation called TESS (TExt Summarization System). The strategy for summarization consists of the following steps

- give a tag to each sentence;
- estimate the importance of each sentence;
- delete the sentences that are not important for the summary;

- produce the summary.

The tags that are given to sentences during the first step contain linguistic informations (e.g. question, opinion, expression of a desire,judgment etc ...) which are mainly determined mainly by the structure of the end of the sentence in Japanese, as well as information about the relationships between sentences (e.g. addition, parallelism, contradiction etc ...) determined using various linguistic information (e.g. anaphora, conjunctions etc ...). The importance of a sentence is quantified from the information written on the corresponding tag, and the system keeps only the sentences that are judged to be important. The selected sentences are finally rearranged in order to obtain a summary in correct natural language (pronoun rearrangement, division of complex sentences etc ...).

4.6 Machine translation (MT)

Machine translation (MT) was historically one of the first tasks that were imagined as application of NLP as early as in the 40's, and remains one that still resists the progress of science. Indeed, current MT systems remain often very poor compared with human translation. In the same time the potential economic rewards linked with the development of MT systems are huge, especially in non English-speaking countries. As a result, Japanese investments in the discipline are colossal and MT softwares from all major computer manufacturers succeed each other very quickly on the shelves in the stores of Akihabara in Tokyo.

4.6.1 NEC's PIVOT

As one of the major computer manufacturer, NEC has developed its own MT system based on an algorithm called PIVOT. Commercialized with the name *Honyaku Adaptator II*, the version for the general public is also based on the pivot method which consists in using *Interlingua*, a formalism for universal semantic representation.

Each sentence to be translated is morphologically, syntactically and semantically analyzed before being represented with the *Interlingua* formalism[2]. Once in this universal formalism, the process is repeated conversely but with a different language.

Acknowledging the imperfection of global translation, the emphasis is put on English assistance for Japanese users through interactive and ergonomic assistance for practical use.

[2]Interlingua represents the concepts linked through 49 possible relationships, e.g. agent-instrument etc ...

4.6.2 NTT's ALT-JE

NTT Communication Science (CS) Laboratory has developed its own MT system named ALT-JE (Automatic Language Translation - Japanese to English), which was still under development in July 1998.

The system carries out a multi-level translation based on a deep morphological analysis of Japanese. The sentence to be translated is separated into a *subjective* part (temporal and modal information) and an *objective* part (the kernel of the sentence). The *objective* part is translated using the multi-level method, where broad rules are applied first (e.g. the transfer of morphological structure tree), followed by idiomatic expressions, structures found in the semantic structure dictionary, and finally default general rules.

ALT-JE uses the concept hierarchy mentioned in the part 3.3.2 as well as a word dictionary, a list of semantic common structures and idioms in a dictionary for the transfer of structures between languages, and a Japanese-English bilingual dictionary that describes syntax and semantics for each word.

The translation software was planned to be commercialized during the fall of 1998.

4.6.3 ATR's MATRIX

The Advanced Telecommunications Research Institute (ATR) developed a MT system too, called MATRIX (Multilingual Automatic TRanslation system for Information eXchange). The special feature of this system is that it was devised for spoken translation of spoken language.

Therefore it includes a real-time speech recognition module, a translation module, and a speech synthesis module including several voices. An experimental system was under development in July 1998, which could recognize 3,000 words and translate 10,000 words. The translation module mixes an example-based and a rule-based approach, and uses a dictionary for the structural transfer between languages. It also uses a series of translation examples that are generalized in the case where there are similarities with the sentence to translate.

4.7 Human/machine interface

One of the challenges of the coming decades will probably be the creation of convivial human/machine interfaces. NLP should play an important role in meeting that challenge in so far as communication in natural language between a machine and a user would represent an important progress toward conviviality. Several experiments are already being made in order to make the machine more "human".

4.7.1 A human/machine dialogue system at NTT

NTT Basic Research Laboratory (BRL) is located in Atsugi in the suburb of Tokyo and was working on the development of dialogue systems between human and machine, including the representation of a face on a computer screen to make it more "human". The main innovation of the system that was under development in July 1998, was the fact that it used the human utterances before the human had finished his sentences to emit sounds, talk or modify its face expression. The communication with the computer thus looks more like a conversation between humans, as it is common for humans to interrupt each other or to approve with small utterances.

4.7.2 The competition DIALEAGUE organized by ETL

The evaluation of automatic dialogue systems is difficult as it is not a problem with one single good answer. Doctor Hasida's team at the Electrotechnical Laboratories (ETL) therefore organizes a competition between different dialogue systems[3]. The goal of the competition is measure the capacity of each system to progress in a dialogue with volunteers through Internet, where the problem is to find a path between two points that would be common to two different subway maps shown to the machine and to the user. The criterion used to judge good dialogues is the its length, i.e. the number of words used.

More than 700 internauts participated in the last version of this tournament in 1997, which enabled a better classification of automatic dialogue systems as well as a classification of users depending on their capacities to test such systems ...

4.7.3 IRENA system at Kyutech

Developed in 1997 for automatic reservation of tickets, IRENA is a dialogue system created by the team of Professor Nomura at the Kyushu Institute of Technology (Kyutech). The system analyses the sentences spoken by the user and generates suitable answers with the aim being ticket reservation. The system therefore has to manage navigation in the dialogue, understand the requests, and look for answers in a database.

Navigation in the dialogue is managed using *dialogue frames*. For request disambiguation the software works in a fuzzy logic framework, for language ambiguities as well as for logic ambiguities, which enables it to generate questions that could make the request more precise. To do this, a fuzzy function is assigned to every linguistic expression (e.g. "about", "from", etc ...) and a fuzzy integration gives the representation of the total request.

The order of expressions, spontaneous utterances, refusals and other linguistic expressions also affect the final fuzzy function.

[3]This competition is described in Japanese at the URL http://www.etl.go.jp/etl/nl/dialeague

Chapter 5

Linguistic resources

This chapter reviews the main linguistic resources (dictionaries and corpus) which were developed in the visited laboratories and which were widely mentioned in the previous chapters.

5.1 EDR's productions

In April 1996, the Japan Electronic Dictionary Research Institute (E.D.R) was created to realize an electronic dictionary for use in advanced NLP research. To build it the company received funds from the Japan Key Technology Center and eight major computer manufacturers : Fujitsu, NEC, Hitachi, Sharp, Toshiba, Oki Electric, Mitsubishi Electric and Matsushita Electric. The project lasted for 9 years, between 1986 and 1994, and led to the creation of five independent dictionaries:

- *Japanese dictionary* It is a 250,000 words dictionary that contains for each word morphological information (pronunciation, accent, etc ...), syntactic information (grammatical characteristics, aspect etc ...) and semantic information (sense explanation and links to all concepts involved).

- *English dictionary* With the same philosophy as the Japanese dictionary, this 190,000 words dictionary defines for every word the concepts that can be attached to it as well as morphological (e.g. inflection, adjacency, pronunciation, accent), syntactic (e.g. POS, countability) and semantic information.

- *Technical dictionary* This dictionary specialized in information processing contains 120,000 Japanese words and 90,000 English words.

- *Concept dictionary* This original dictionary describes and classifies the set of 400,000 concepts needed to fully understand the meaning of each word. The classification is based on supra/super relationships. The

description contains binary semantic relationships between concepts (e.g. agent/action or object/action)

- *Bilingual dictionary*
- *Co-occurrence* This is a table that contains information about the possibility of word sequences inside sentences, and about concept collocations.
- *Japanese and English corpus* This corpus is made of 220,000 Japanese and 160,000 English sentences. Each sentence is morphologically, syntactically and semantically parsed.

These dictionaries have been used to develop the morphological parser JUMAN at Kyoto University. Since 1996, E.D.R. has joined the ANSI Ad-Hoc Group for Ontology Standards and is trying to link its dictionaries with Wordnet.

In July, 1998, the cost to buy the dictionaries was 100,000 JPY (around 1,000 USD) for universities and 1,200,000 JPY (around 12,000 USD) for companies.

5.2 NTT's productions

This telecommunication giant has produced dictionaries and corpus for its research in NLP, especially machine translation. These resources were later used by many research centers.

NTT's dictionary is a 400,000 words dictionary. For each word, the pronunciation and the canonical form as well as syntactic and semantic information are provided. Semantic information is based on a 3,000 semantic attributes hierarchical graph that uses "is a" and "has a" relationships between concepts. The semantic attributes are provided for every word of the dictionary.

Parallel to the Japanese dictionary NTT developed a Japanese/English bilingual dictionary for classical structures and idioms, with 17,000 entries, including 6,000 ambiguous verbs. This dictionary contains the equivalences between Japanese and English structures.

Finally, a Japanese-to-English dictionary containing syntactic and semantic information is also available.

5.3 Kyoto University corpus project

A corpus project was under development at Kyoto University, whose goal was to create the corpus semi-automatically, to provide grammatically parsed sentences and to improve the automatic parsers at the same time.

The corpus contained 20,000 sentences in July 1998. Sentences were automatically parsed with JUMAN for the morphology and KNP for the syntax. Every sentence was then checked and eventually modified by humans, and the errors are used to improve the parsing algorithms used. The rate of growth of the corpus was of about 40 sentences per hour and per person.

This corpus can be downloaded through Internet on Kyoto University web site, but it is necessary to buy CD-Roms of the newspaper that was used to provide the sentences.

5.4 ATR's resources

The Advanced Telecommunications Research Laboratories (ATR) have developed two types of resources for research in NLP. One should notice that both of them are in English

- *Hierarchical clustering of words* Starting from a database of English texts (the Wall Street Journal), researchers from ATR have created a classification of the 70,000 most used words into classes hierarchically sorted. The result is a binary tree where each node is a class. The tree was built automatically and iteratively, mainly by clustering together the classes that were often used in similar situations. The algorithm uses the mutual information between classes defined by:

$$MI(a,b) = P(a,b)\log\frac{P(a,b)}{P(a)P(b)}$$

  The resulting binary tree where each node can be coded with a series of bytes is a semantic organization that can be used for various applications

- *ATR/Lancaster Treebank* This 730,000 words corpus is divided into 950 documents with lengths ranging from 30 to 950 words. Texts were chosen so as to maximise diversity of text lengths, topics and author. All texts are tagged with the tagging system used at ATR. Sentences are parsed with respect to the ATR English grammar, which is a feature-based context-free phrase-structure grammar that contains 67 features and 1,100 rules.

5.5 ETL's GDA project

The Electrotechnical Laboratory (ETL) in Tsukuba, which developed the multilingual *MULE* environment that is available on the GNU Emacs v.20, was trying to promote an annotation standard for HTML documents published on Internet : the Global Document Annotation (GDA). This standard enables computers to recognize semantic and pragmatic structures of HTML

documents. The initiators of that project hope that a huge quantity of tagged documents will appear on the Internet, which could in particular be used as a linguistic corpus. To promote this standard, a collection of tags was proposed to enable computers to guess the structures of a document, and several applications were developed like automatic translation of GDA-tagged document, data-mining, automatic summarization or automatic design of slides for a presentation concerning such a document.

Chapter 6

Laboratories

Last but not least, this chapter gives a detailed list of the laboratories that contributed to that report by opening their doors and showing some of their research topics

6.1 Advanced Telecommunications Research Institute International (ATR)

Coordinates
ATR Interpreting Telecommunications Research Laboratories

Internet site

`http://www.itl.atr.co.jp/`

Some research topics
Machine translation, parser, tree corpus.

6.2 Electrotechnical Laboratory (ETL)

Coordinates
Ministry of International Trade and Industry (MITI)
Electrotechnical Laboratory
Natural Language Learning Laboratory

Internet site

`http://www.etl.go.jp`

Some research topics
Dialogue system, semantic representation, GDA annotation standard.

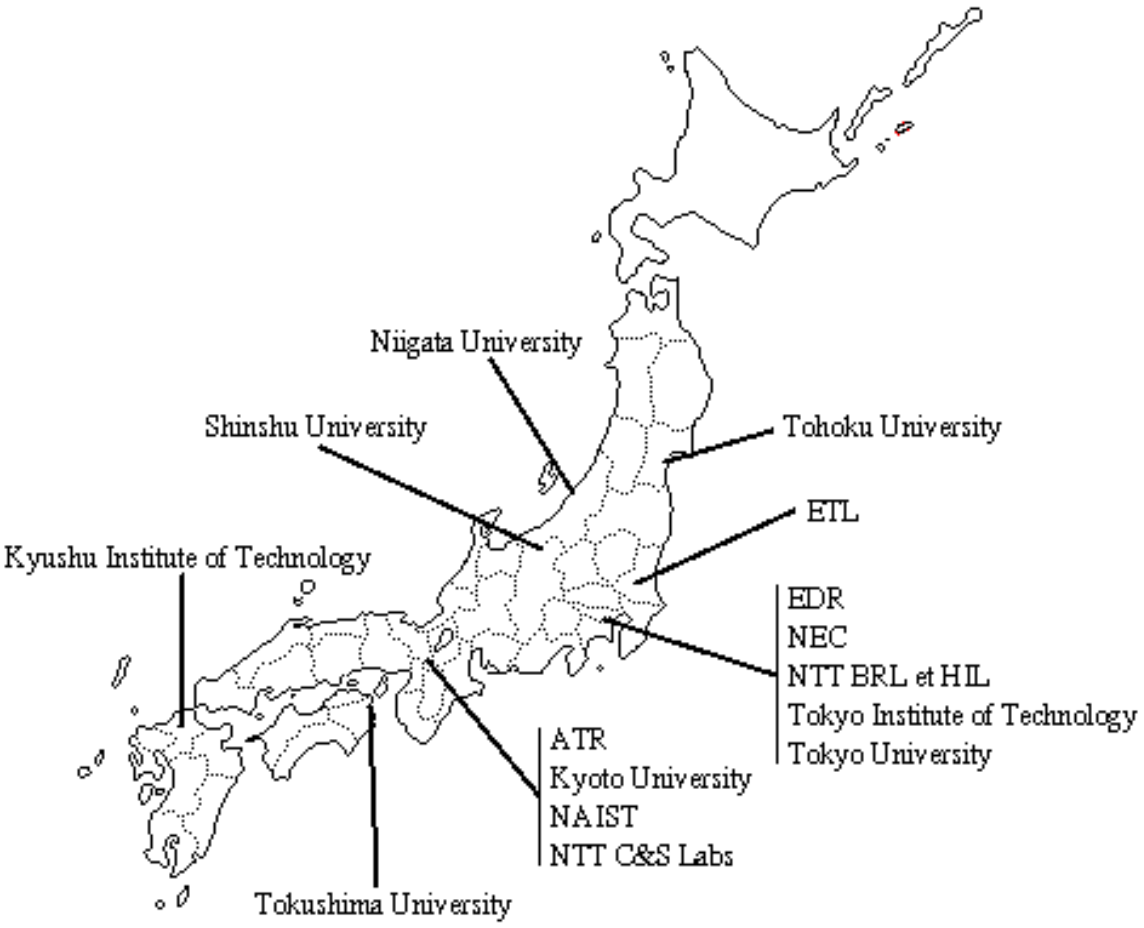

Figure 6.1: Laboratories

6.3 Japan Electronic Dictionary Research Institute, Ltd (EDR)

Coordinates

Japan Electronic Dictionary Research Institute, Ltd

Internet site

`http://www.iijnet.or.jp/edr/`

Some research topics

Dictionaries, semantic representation and corpus.

6.4 Kyoto University

Coordinates

Language Media Laboratory (Nagao/Kurohashi Lab)
School of Electrical and Electronic Engineering
Faculty of Engineering
Kyoto University

Internet site

`http://www-nagao.kuee.kyoto-u.ac.jp/index-e.html`

Some research topics

Morpho-syntactic parser, text generation, statistics for text analysis and information retrieval.

6.5 Kyushu Institute of Technology (Kyutech)

Coordinates

Nomura and Nakamura Laboratory
Department of Artificial Intelligence
Faculty of Information Engineering
Kyushu Institute of Technology

Internet site

`http://www.dumbo.ai.kyutech.ac.jp/htdocs/nomura-ken/nomura-ken-e.html`

Some research topics

Information extraction, summarization, dialogue system.

6.6 Nara Institute of Science and Technology (NAIST)

Coordinates

Matsumoto Laboratory (Computational Linguistic Laboratory)
Graduate School of Information Science
Nara Institute of Science and Technology

Internet site

`http://cactus.aist-nara.ac.jp/lab-english/home-e.html`

Some research topics

Syntax and semantics of natural languages, discourses.

6.7 NEC

Coordinates

Pattern Analysis Technology Group
C&C Media Research Laboratories

Internet site

`http://www.nec.co.jp`

Some research topics

Machine translation, information retrieval.

6.8 Niigata University

Coordinates

Miyazaki Laboratory (NLP Laboratory)
Department of Information Engineering
Faculty of Engineering
Niigata University

Internet site

`http://www.nlp.info.eng.niigata-u.ac.jp/nlp/index.html`

Some research topics

Morpho-syntactic parsing.

6.9 NTT Basic Research Laboratories (BRL)

Coordinates

Information Science Research Laboratory
NTT Basic Research Laboratories

Internet site

`http://www.brl.ntt.co.jp/info/index.html`

Some research topics

Dialog, human/machine interfaces.

6.10 NTT Communication Science Laboratories

Coordinates

NTT Communication Science Laboratories

Internet site

`http://www.kecl.ntt.co.jp`

Some research topics

Machine translation, information retrieval, semantic representation, dictionaries.

6.11 NTT Human Interface Laboratories

Coordinates

NTT Human Interface Laboratories

Internet site

`http://www.hil.ntt.co.jp/top/index_e.html`

Some research topics

Dialogue, information extraction.

6.12 NTT Software Laboratories

Coordinates

Global Computing Laboratory
NTT Software Laboratories

Internet site

`http://www.kokono.net`

Some research topics
Mobile information search.

6.13 Shinshu University

Coordinates
Nakano and Murayama Laboratory
Department of Information Engineering
Faculty of Engineering
Shinshu University

Internet site

`http://sunak2.cs.shinshu-u.ac.jp/index.html`

Some research topics
Analysis and recognition of written documents.

6.14 Tohoku University

Coordinates
Aso Laboratory
Department of Electrical and Communication Engineering
Graduate School of Engineering
Tohoku University

Internet site

`http://www.aso.ecei.tohoku.ac.jp/index-e.html`

Some research topics
Written documents analysis, characters recognition.

6.15 Tokushima University

Coordinates
Aoe Laboratory
Department of Information Science and Intelligent Systems
Faculty of Engineering
Tokushima University

Internet site

`http://www-b3.is.tokushima-u.ac.jp/aoe/index.html`

Some research topics

Automatic classification, information retrieval.

6.16 Tokyo Institute of Technology

Coordinates

Tanaka and Tokunaga Laboratory
Department of Computer Science
Graduate School of Science and Engineering
Tokyo Institute of Technology

Internet site

`http://tanaka-www.cs.titech.ac.jp/tanaka-home-e.html`

Some research topics

Morpho-syntactic parsing, information retrieval.

6.17 Tokyo University

Coordinates

Tsujii Laboratory
Department of Information Science
Faculty of Science
Tokyo University

Internet site

`http://www.is.s.u-tokyo.ac.jp/~tsujiilab/`

Some research topics

Morpho-syntactic parsing.