様式例：課程博士用

19年　6月　27日

電子。情報工学専攻	学籍番号	049302
申請者氏名	Ayu Purwarianti	

指導教員氏名	中川聖一

論　文　要　旨(博士)

論文題目	Developing Cross Language Systems for Language Pair with Limited Resource – Indonesian-Japanese CLIR and CLQA – (言語資源が希少な言語対を対象とする言語横断処理システムの開発 – インドネシア語-日本語 CLIR および CLQA –)

(要旨 1,200字程度)

Researches on cross language text processing systems have become an interesting research area, including CLIR (Cross Language Information Retrieval) and CLQA (Cross Language Question Answering) systems. For major languages, there are various available resources such as a parallel corpus, a rich bilingual dictionary, a high performance machine translation software, etc. In order to translate an English sentence into Japanese, one can use the available free machine translation such as Babelfish, Excite, etc. But this is not always the case, especially for minor languages such as Indonesian. For Indonesian, until now, obtaining a rich resource for the translation is quite impossible. To have resources such as the major languages do, one has to spend a lot of work hours.

In this thesis, I deal with the cross language systems for Indonesian, a language with limited resources. I developed some systems for Indonesian language such as Indonesian-Japanese CLIR, Indonesian monolingual QA, Indonesian-English CLQA and Indonesian-Japanese CLQA. The main aim of these researches is to propose methods to handle the limited resource problem.

In the Indonesian-Japanese CLIR, I propose a query transitive translation system of a CLIR or a language pair with limited data resources. The method is to do the transitive translation with a minimum data resource of the source language (Indonesian) and exploit the data resource of the target language (Japanese). There are two kinds of translation, a pure transitive translation and a combination of direct and transitive translations. In the transitive translation, English is used as the pivot language. The translation consists of two main steps. The first is a keyword translation process which attempts to make translation based on available resources. The keyword translation process involves many target language resources such as the Japanese proper name dictionary and English-Japanese (pivot-target language) bilingual dictionary. The second step is a process to select some of the best available translations. The mutual information score (computed from target language corpus) is combined with the TF×IDF score in order to select the best translation. The result on NTCIR 3 (NII-NACSIS Test Collection for IR Systems) Web Retrieval Task showed that the translation method achieved a higher IR score than the machine translation (using Kataku (Indonesian-English) and Babelfish/Excite (English-Japanese) engines). The performance of transitive translation was about 38% of the monolingual retrieval, and the combination of direct and transitive translation achieved about 49% of the monolingual retrieval which is comparable to the English-Japanese IR task.

In the monolingual Question Answering (QA) system, I have developed a QA system for a limited resource language (e.g. Indonesian) which employs a machine learning approach. The QA system consists of two key components: "question classifier" and "answer finder", which are based on Support Vector Machines (SVM). I also developed some supporting tools such as an easily built POS tagger and a shallow parser for the question. These supporting tools are built with small human efforts. In the development, there are 3000 questions for 6 answer types, collected from 18 Indonesian. For the evaluation data, there are 71,109 Indonesian news articles available on Web. In the experiments, some feature combinations for the SVM were compared. All features used are extracted from the available language resources. One of the important features is a bi-gram frequency between the intended word and some defined words. This feature is introduced to cope with the resource poorness. For the question classification task, the system achieved about 96% accuracy. The answer finder achieved MRR of 0.52 on the first answer as the exact correct answer. Using this machine learning approach, I argue that this monolingual QA system can be adapted easily to other limited resource language.

For the CLQA research, I adopted the approach used in the Indonesian monolingual QA into the Indonesian-English CLQA system. The Indonesian-English CLQA system was built from Indonesian question analyzer system, Indonesian-English translation using a bilingual dictionary, English passage retriever and English answer finder. Different with other Indonesian-English CLQA systems, I did a bilingual dictionary translation in order to make the keyword coverage larger. The translation module is equipped with a transformation module for Indonesian borrowed words such as "prefektur"(from "prefecture"), "Rusia"(from "Russian"), etc. The translation results are combined into a boolean query to retrieve relevant English passages. Features of translated question keywords and passages are used to define the answer in the English passages. The bi-gram frequency feature used in the Indonesian answer finder for each word in the passage is replaced by the WordNet distance feature. This replacement is done easily without adding any mapping tables. In the experiments, 2553 questions were used as the training data and 284 questions were used as the test data. These questions were collected from 18 Indonesian college students. Using this in-house data, the question answering achieved the accuracy of 25% for the first correct answer. Experiments were also conducted using the translated test questions from NTCIR 2005 CLQA data. For the NTCIR 2005 CLQA data, my Indonesian-English CLQA system is superior to others except for one with a rich translation resource. I also did an experiment for various sizes of training data which shows that the size of training data does influence the accuracy of a CLQA system.

In the Indonesian-Japanese CLQA, I used a transitive approach in translation and passage retrieval phase. Similar with the Indonesian-Japanese CLIR, English is used as the pivot language in the transitive translation with bilingual dictionaries. The experiment shows that the passage retriever for transitive translation using mutual information score and TF×IDF score as the translation filtering could enhance the performance to be higher than the direct translation. Furthermore, using English passage retriever result as the input for the Japanese passage retriever gives much higher passage retrieval performance compared to the one with only query as the input. The answer finder process employs easily gained features including the POS information yielded by Chasen (an available Japanese morphological analyzer). Even though the Indonesian-Japanese question answering performance is lower than the Indonesian-English CLQA but it is higher than other research using a similar technique which employs text chunking process in an English-Japanese CLQA.