Masters Thesis Proposal: Implementation of English-to-Japanese Machine Translation of Personal Names for Cross-Lingual Information Retrieval in the Context of a Question Answering System

Scott Judy

Language Technologies Institute,
School of Computer Science,
Carnegie Mellon University

Thesis Committee Members:
Robert Frederking
Teruko Mitamura
Eric Nyberg

Oct 30, 2003

1 Introduction

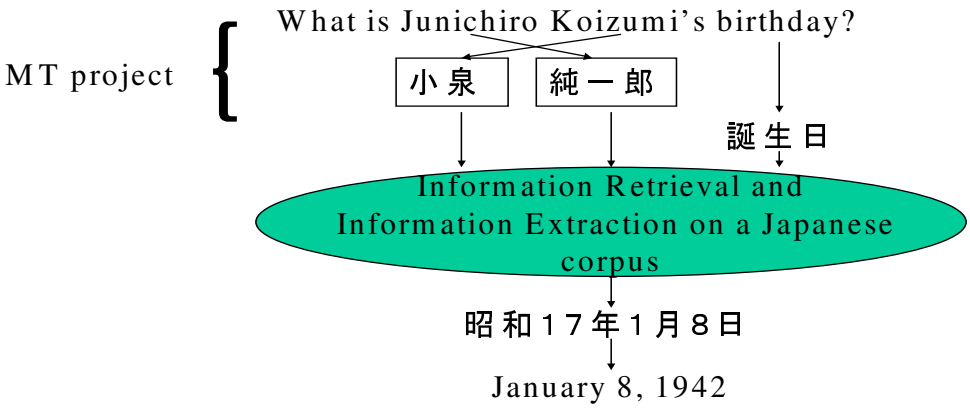
This thesis proposal describes the need for and suggests the implementation of a new method of translating personal names from English and Japanese in the context of an open-domain question answering system. It is composed of several sections: problem background, expected contribution of the thesis work, previous findings, hypothesis, design of the experimental system, and evaluation.

2 Background: Keyword Translation in a Multilingual Question Answering System

The JAVELIN system is an ARDA-funded research project to explore open-domain question answering[EN1][EN2]. It employs a modular architecture making it possible to incorporate many different tools and approaches to extracting answers questions from multiple corpora.

One of the things that makes the JAVELIN project especially challenging is that it attempts to do open-domain question answering. This means that there is no guarantee that the system has any specialized knowledge about the domain of the question being asked, other than the corpora used, and that even in that situation it must still function well enough to produce a "best-guess" answer and guide the analyst toward a valid conclusion.

However, there is an additional requirement of JAVELIN – multilingual capability – that presents a difficulty. Multilingual capability means that an English question's answer may be found in a Japanese corpus. But this requires the translation of questions or at least question keywords into Japanese. An illustration of retrieving the answer to the English question "What is Junichiro Koizumi's Birthday?" from a Japanese corpus is given here.



While question translation can be done to some extent with Machine Translation (MT) systems, bilingual dictionaries, or parallel corpora, these methods do not degrade gracefully for a word that is not covered by the bilingual information source. They simply fail. This may be acceptable for one or two words in a question, but it presents a special problem if people's names cannot be translated, because names are a crucial component of any question in which they appear. For example, the correct answer to the question above cannot be obtained without a correct translation of the name "Junichiro Koizumi" to "小泉純一郎". A method of translating names that degrades more gradually is called for.

2.1 E→J Translation of Names in Commercial MT Systems

Most of the commercial machine translation systems available for English and Japanese deal with the translation of names in a dictionary lookup fashion, essentially treating names the same as any other word. If a name can be found in the lexicon, it is translated, and if not, then it is simply left in the orthography of the original language.

This was confirmed by analyzing two of the more well known Japanese to English web translation services, Systran and Amikai.. Both produced very similar results, though Amikai seemed to have better name coverage and produce more grammatical sentences overall. The example sentences below are the results of translations run on Amikai.

2.1.1 English Names

At first glance, Amikai appears to deal with English names very well, and the sentences pairs below both translate correctly in both directions:[1]

George Bush is famous. ↔ ジョージ・ブッシュは有名です。

Scott Judy is not famous. ↔ スコット・ジュディは有名ではありません。

But on closer inspection some problems appear. For example Amikai produces the following reasonable English to Japanese translation:

Erik Barnett is not famous. → エリック・バーネットは有名ではありません。

However, the back translation ends up as a plausible, but different name:

Eric Burnet is not famous. ← エリック・バーネットは有名ではありません。

Because of the limited repertoire of phonemes in the Japanese language and the plurality of spellings in the English language, the mapping is in fact ambiguous, so the system perhaps cannot be faulted for this performance unless the user is willing to accept a ranked list as their translation output.

However, with the next sentence/translation pair it becomes clear that if a name is not found in the system's lexicon, the translation fails completely:

[1] The corresponding names in the English and Japanese sentences are underlined. The arrows between the sentences indicate in which direction the translation works.

Cindi is not famous. ↔ Cindi は有名ではありません。

In contrast, the following sentence, with a known spelling, works as expected:

Cindy is not famous. ↔ シンディーは有名ではありません。

2.1.2 Japanese Names

This issue becomes more noticeable for E→J translation in the case of Japanese names, which the system seems to handle differently, though the approach is still lexicon-based. The following sentence works fine:

Junichiro Koizumi is famous. ↔ 小泉純一郎は有名です。

But Junichiro Koizumi is the prime minister of Japan, and it is not surprising that the system can translate his name correctly. The following is the result of trying to translate a more obscure Japanese name:

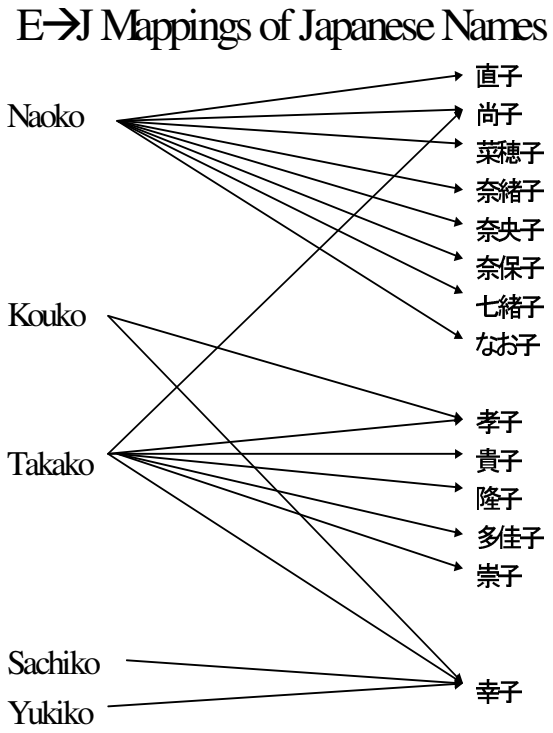
Mr. Susumu Sato is not famous. → Susumu 佐藤氏は有名ではありません。

Mr. Susumu Sato is not famous. ← 佐藤進氏は有名ではありません。

The translation from English to Japanese and Japanese to English succeeds for "Sato" because the name is unambiguous, having only one representation in Kanji. Neither Amikai nor Systran attempts to translate the given name "Susumu" into Japanese, probably because there are several ways it could be written and the designers judged it better to preserve the ambiguity than choose a possibly incorrect name. Both systems,

however, attempt to translate the given name "進" into English, Amikai choosing the more likely "Susumu", and Systran choosing "Shin", another possible reading.[2]

As the following chart illustrates, the mapping between English versions of Japanese names and the Japanese names themselves is a complex, many-to-many mapping, though in general the mapping from Japanese to English is less ambiguous than the other direction.



Because of this inherent ambiguity, any question answering system that simply searches for the one "right" Japanese version of a name is not likely to produce great results.

[2] Perhaps this behavior comes from the consideration by designers that Japanese readers can understand Japanese names in English, while monolingual English readers cannot get any information from a Japanese name in Japanese.

There is one other issue in E→J name translation: E→ J conversion of Japanese and Chinese[3] names usually involves a conversion into Chinese characters, while conversion of other names, such as western names, involves a transliteration into the phonetic katakana characters. This implies that any system that will attempt to translate a name must be able to distinguish between the languages of origin of the name.

2.1.3 Conclusions About Current MT Systems

Both the Amikai and Systran systems exemplify the strengths and weaknesses of lexicon-based translation of names. When the name is present in the lexicon and unambiguous, the translation succeeds perfectly, but when the name is not present or there is ambiguity, translation is either not attempted, or only one candidate of several is output. Such a system cannot be expected to perform adequately as a component of an open-domain question answering system.

Two enhancements are required to make such a system useful for translating queries for an open-domain question answering system:

- The ability to attempt translation of a name by using the characters making up the name itself and rules about the composition of names is required. This is analogous to the ability to apply lexical knowledge and grammar rules for machine translation of a sentence when the whole sentence is not found in a translation memory.
- The ability to produce a ranked list of candidates is also desirable, since we may not be sure of which candidate is correct, but a good answer may be found by combining multiple candidates with the other question keywords used in the information retrieval module of the question answering system.

[3] It is this author's intuition that Chinese names should be treated as a distinct category for conversion from English to Japanese, and that what is called for is essentially conversion from English to Chinese, followed by a conversion of character mappings into Japanese. While that would be an interesting eventual addition to the system proposed, it is beyond the scope of this thesis.

3 Previous Work

A review of machine translation literature did not address the issue of how to translate names between languages with different orthography. It is only when looking at Cross Lingual Information Retrieval (CLIR) literature that some answers could be found. It has been shown, for instance, that a bilingual lexicon, when used for CLIR, results in substantially worse performance.[YY] However, more importantly, this approach is not tractable for proper names, because very large bilingual lexicons with good coverage of names simply do not exist.[JN][YY]

The CLIR literature indicates that bilingual lexicons derived from sentence-aligned parallel corpora perform much better, often as well or better than monolingual IR systems[YY], implying that they can accurately identify correlations between words, including names. However, these corpora cannot be expected to contain all of the names one might need to deal with in a CLIR task. In fact, many names are probably not in any English news corpus until an individual with that name starts making news (e.g. "Osama bin-Laden"). Since a question answering system, especially one designed for analysts, must deal with information sources not in the public domain, it is to be expected that obscure names not available in any freely available training corpora will be plentiful.

Clearly a method for attempting a translation of a name even when we do not have bilingual data for that particular name is needed, and a search for such work found work on the transliteration of names between English and the following languages: Japanese[AF][EB][JH][KK][YA2], Chinese[AF][WL][YA2], Arabic[YA1][YA2], and Hindi[AN]. More similar work likely exists. All of this research, however, makes the assumption that the original language of a name is known, and that it is desired only to convert a name between English and its original language.

Further, no research was discovered on the problem of translating Japanese names from English into Japanese. This may be because in general the problem is intractable, with

many possible candidates. This thesis will argue that a task-based translation for the purpose of information retrieval is, however, possible.

Finally, no work describing an attempt to distinguish between Japanese and other names appearing in an English text was found, though it is a necessary step in translation.

4 Expected Contribution

Names, especially Japanese names, are impossible to list exhaustively in a dictionary to a satisfactory degree, and translation of names is crucial. However, MT systems currently treat names the same as any other vocabulary and attempt to look them up. When this fails, the name may not be translated at all, but simply copied. Even if the lookup succeeds, the name returned may not be correct for the specific query, since there is ambiguity in the mapping between names in English and Japanese.

The system proposed for implementation would demonstrate a framework in which a system may choose a strategy most appropriate to translation of a particular name. This is similar in concept to Multi-Engine Machine Translation (MEMT), except that it operates at the word level instead of a sentence level, and allows different approaches to producing a translation of the word, the outputs of which can be combined.

Work has been done on transliteration between languages, but it has not been incorporated into a system that can choose this as one of several possible strategies. The proposed system would differentiate between English and Japanese names, attempting one type of transliteration of English names to katakana, while choosing a different approach with Japanese names, which need to be represented in kanji.

Previous work using parallel corpora in CLIR to find associations is related, but this provides no way to guess about names which are not in the parallel corpora. The proposed system could incorporate such correlations as one source of information, but could also fall back to other types of transliteration when it is not available.

The combination of several different methods for a multi-engine approach to translation of names has one other advantage for question answering and information retrieval applications – the ability to produce a ranked list of names would allow multiple candidates to be placed into the search terms passed to the information retrieval engine.

5 Hypothesis

The hypothesis of the proposed thesis is that machine translation of personal names from English to Japanese in the context of information retrieval in a cross-lingual question answering system need not rely soley on bilingual dictionaries and parallel corpora, but can be improved by detecting the language of origin (e.g. Japanese or English) of each name in an English sentence and applying a translation strategy appropriate to that type of name, and that the problem of many-to-many mapping of names in English and Japanese can be mitigated by the information retrieval task, because of the presence of other translated keywords from the sentence.

6 System Design

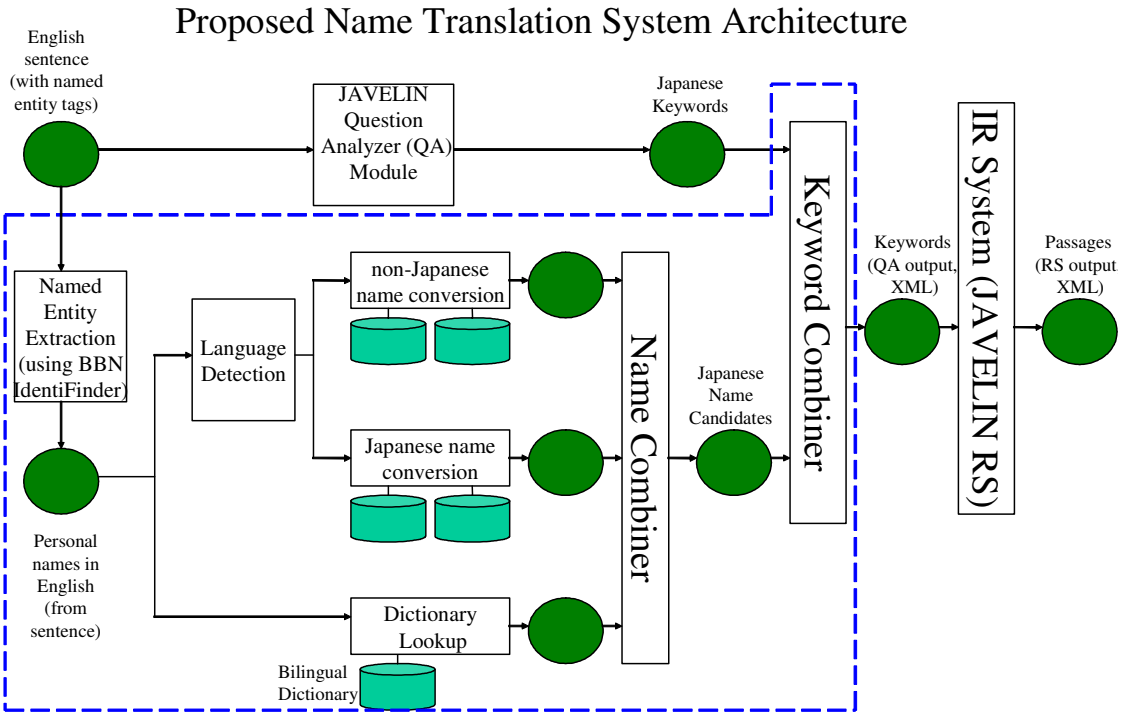
The JAVELIN system performs question answering, but this thesis will focus narrowly on the subtask of providing better Japanese translations of names in order to retrieve more relevant documents upon which to perform information extraction, in effect improving the performance of the IR sub-system.

The system described here will be designed to function as part of the JAVELIN system. It will be implemented in two parts:

- an addition to the "Question Analyzer" module of the system – the module that takes as input an English Question and provides as output keywords for the information retrieval module (the "Retrieval Strategist") to use in a variety of languages.

- an expansion of the index of the Japanese corpus to include possible readings (in hiragana) of personal names that are written in Chinese characters.

The following chart illustrates the architecture of the system proposed and how it will interface with the JAVELIN Question Analyzer module.



This chart shows the insertion of the name translation system described in this thesis into the JAVELIN Question Analyzer (QA) module.

The items to be built for this thesis are enclosed in the dotted line.

The area enclosed in the dashed line indicates the scope of the changes to the JAVELIN Question Analzyer module that will be implemented for this thesis. The information from the corpus that is indexed by the information retrieval module (the "Retrieval Strategist") will also be expanded to include possible readings of the personal names. This is possible because the Japanese corpus has been preprocessed with a combination of the Chasen[YM] and YamCha[TK] systems, and the named entities are already tagged. (The Japanese corpus used in Javelin consists of the full text of the Mainichi Shinbun, a Japanese newspaper, for the years 1998-1999.)

1. The system will accept English questions as input and extracts a list of personal names identified by BBN IdentiFinder.
2. These names will then be classified as either Japanese or non-Japanese in origin, based on the use of regular expressions to identify names that may be sequences of Japanese syllables.
3. If a name is classified as Japanese in origin, it is transliterated from English into Japanese hiragana by treating the string as a sequence of Japanese syllables in "romaji" (Roman characters). [4]
4. If the name was classified as non-Japanese in origin, transliteration will be attempted using a cascaded finite-state automaton system based on the Japanese-to-English transliteration system developed by Kevin Knight [KK].
5. Independent of classification and language-of-origin-dependent strategies, a lookup of each name in a bilingual dictionary (ENAMDIC[EN3] will be attempted.
6. The name combiner will then combine all of the name candidates produced by each module into a list of name candidates to be queried.
7. The keyword combiner will then combine the name candidates with the keyword candidates produced by the existing question analyzer module. It is the keyword combiner's output that will server as a query to the information retrieval system.

In addition to these changes, the index of the corpus will also be changed in order to index possible readings of each Japanese (kanji) name in hiragana, based on the possible readings of the constituent characters of the name. The queries will be weighted so that a match between the hiragana produced by the Japanese name conversion module (in the QA module) and the hiragana added to the index for a kanji name can be weighted differently than a direct match between the original kanji in the corpus and the kanji produced by the expanded QA module looking up the name in the EDICT bilingual dictionary.

[4] Where this conversion is ambiguous, as with long and short vowels, both candidates will be produced. (e.g. Since "Ono" may be "大野" or "小野", both "おおの" and "おの" would be produced.)

7 Evaluation

It is hoped that the system described in this proposal will result in more relevant documents being returned than when using the standard JAVELIN "Question Analyzer" module (with web-based MT software and bilingual dictionary lookup only) to produce keywords for Japanese document retrieval, and in a higher percentage of the documents returned being relevant, where a document's relevance to a question is determined by whether or not it contains the answer to that question.

Evaluation will be performed as follows: A test suite of TREC-type questions containing personal names, some Japanese and some non-Japanese, will be used as input to the JAVELIN system. These questions will be designed such that their answers are contained in the Mainichi Shinbun corpus used by the system, except for a small number of questions that will have no answer in the corpus.

Both the regular JAVELIN system (without the personal name translation enhancements described in this proposal) and the enhanced JAVELIN system (with the enhancements) will be run on the questions in the test suite, and the following items will be determined by human evaluation:

- The number of documents returned by each system that contain answers to each question.
- The number of documents returned by each system that do not contain answers to each question.
- The total number of documents returned by each system for each question.

We desire to find the increase, if any, in the number of relevant documents gained by the name translation system (recall), and we also desire to determine the change in the number of relevant documents (i.e. those containing the answer) as a percentage of the total documents returned (precision).

The precision of both systems can be evaluated and compared. However, because there will be no canonical list of relevant corpus documents available for each question, the recall will be evaluated only as a change in the absolute number of relevant documents found from one system to the other, and not as a change in the percentage of all of the relevant documents in the corpus.

8 References

[AF] Atsushi Fujii and Tetsuya Ishikawa. Japanese/English Cross-Language Information Retrieval: Exploration of Query Translation and Transliteration. Computers and the Humanities, Vol.35, No.4, pp.389-420, Nov. 2001.

[AL] Ariadna Font Llitjós, "Improving Pronunciation Accuracy of Proper Names with Language Origin Classes", Carnegie Mellon University, Language Technologies Institute, Master's Thesis, 2001.

[AN] A. Natrajan, A. L. Powell, and J. C. French "Using N-grams to Process Hindi Queries with Transliteration Variations" Technical Report CS-97-17, Dept. of Computer Science, Univ. of Virginia, July 1997.

[CL] Chen, Hsin-Hsi and Jen-Chang Lee (1996): Identification and classification of proper nouns in Chinese Texts. *Proceedings of COLING-96*, Vol. 1, pp. 222-229, Copenhagen, Denmark.

[EB] Eric Brill, Gary Kacmarcik and Chris Brockett. "Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs," NLPRS 2001.

[EN1] Nyberg E. and T. Mitamura, "Evaluating QA Systems on Multiple Dimensions," Proceedings of LREC 2002 Workshop on QA Strategy and Resources, May 28th, Las Palmas, Gran Canaria.

[EN2] Nyberg, E., T. Mitamura, J. Carbonell, J. Callan, K. Collins-Thompson, K. Czuba, M. Duggan, L. Hiyakumoto, N. Hu, Y. Huang, J. Ko, L. Lita, S. Murtagh, V. Pedro and D. Svoboda, "The JAVELIN Question-Answering System at TREC 2002", unpublished manuscript, November 2002.

[FG] Gey, Fredric C. "Research to Improve Cross-Language Retrieval - Position Paper at CLEF", presented at the Cross-Language Evaluation Forum, Lisbon Portugal September 21-22, 2000, to be published by Springer, 2001.

[JH] Jack Halpbern, "Lexicon-Based Orthographic Disambiguation in CJK Intelligent Information Retrieval", COLING 2002.

[JN] Jian-Yun Nie, Michel Simard, Pierre Isabelle, Richard Durand: Cross-Language Information Retrieval Based on Parallel Texts and Automatic Mining of Parallel Texts from the Web. SIGIR 1999: 74-81.

[YM] Yuji Matsumoto, Akira Kitauchi, Tatsuo Yamashita, Yoshitaka Hirano, Hiroshi Matsuda, Kazuma Takaoka, Masayuki Asahara. "Japanese Morphological Analysis System ChaSen version 2.2.1" User's Manual, Dec, 2000. (Available at http://chasen.aist-nara.ac.jp/chasen/doc/chasen-2.2.1.pdf)

[LB] Ballesteros, L. and Croft, W. B. (1997). Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval. In Proceedings of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA.

[LC] Wei-Hao Lin and Hsin-Hsi Chen, "Backward Machine Transliteration by Learning Phonetic Similarity", *Sixth Conference on Natural Language Learning*, Taipei, Taiwan, August 31 - September 1, 2002

[KK] K. Knight and J. Graehl. "Machine Transliteration," Computational Linguistics, 24(4), 1998.

[NY] Yegao, Ning and Ning Yun. 2000. Chinese Personal Names. Singapore: South East Printing Pte Ltd, Singapore.

[PO] P.G. O'Neill. 1993. Japanese Names: A Comprehensive Index by Characters and Readings. New York and Tokyo: Weatherhill, Inc.

[TK] Taku Kudo, Yuji Matsumoto. Chunking with Support Vector Machines, NAACL 2001. (Also see http://cl.aist-nara.ac.jp/~taku-ku/software/yamcha/ for general information on Yamcha)

[TS] T. Schultz, I. Rogina, A. Waibel, "LVCSR-based language identification." Proc. ICASSP-96, Atlanta, USA, 1996.

[VP] V. Pagel, K. Lenzo, and A. Black. 1998. "Letter to sound rules for accented lexicon compression." *Proceedings of the 1998 International Conference on Spoken Language Processing*, Sydney, Australia.

[WL] Wei-Hao Lin and Hsin-Hsi Chen, "Backward Machine Transliteration by Learning Phonetic Similarity", Sixth Conference on Natural Language Learning, Taipei, Taiwan, August 31 - September 1, 2002.

[YA1] Y. Al-Onaizan and K. Knight. "Machine Transliteration of Names in Arabic Text" Proc. of ACL Workshop on Computational Approaches to Semitic Languages, 2002.

[YA2] Y. Al-Onaizan and K. Knight. "Translating Named Entities Using Monolingual and Bilingual Resources", Proc. of the Conference of the Association for Computational Linguistics (ACL), 2002.

[YY] Y. Yang, J. Carbonell, R. Brown, and R. Frederking. Translingual information retrieval: Learning from bilingual corpora. In Artificial Intelligence Journal special issue: Best of IJCAI-97, 1997

Web Sites Referenced in this Proposal

[AM] Amikai website: http://www.amikai.co.jp

[EN3] ENAMDICT downloadable from: http://ftp.cc.monash.edu.au/pub/nihongo

[SY] Systran website: http://www.systransoft.com