Extraction of Lexical Translations from Non-Aligned Corpora

Kumiko TANAKA
Faculty of Engineering
The University of Tokyo
7–3–1 Hongo, Bunkyo-ku
Tokyo 113 JAPAN
kumiko@ipl.t.u-tokyo.ac.jp

Hideya IWASAKI*
Educational Computer Centre
The University of Tokyo
2–11–16 Yayoi, Bunkyo-ku
Tokyo 113 JAPAN
iwasaki@rds.ecc.u-tokyo.ac.jp

Abstract

A method for extracting lexical translations from non-aligned corpora is proposed to cope with the unavailability of large aligned corpus. The assumption that "translations of two co-occurring words in a source language also co-occur in the target language" is adopted and represented in the stochastic matrix formulation. The translation matrix provides the co-occurring information translated from the source into the target. This translated co-occurring information should resemble that of the original in the target when the ambiguity of the translational relation is resolved. An algorithm to obtain the best translation matrix is introduced. Some experiments were performed to evaluate the effectiveness of the ambiguity resolution and the refinement of the dictionary.

1 Introduction

Alignment of corpora is now being actively studied to support example-based automatic translation and dictionary refinement. Focusing on the latter, in order to obtain lexical translations, the maximum likelihood method is applied to roughly aligned corpus. One of the problems of this method is that it needs a large amount of aligned corpus for training (Brown, 1993).

When it exists, a qualified dictionary is also likely to exist, because it should have been created and used when the corpus in the source language was translated by hand to make the aligned corpus. There are few requirements to improve dictionaries in such a case. On the other hand, when a large amount of aligned corpus does not exist but only two independent corpora do, for example, the corpora between two 'not so international' languages or those in a constrained domain, the low quality dictionaries need to be improved.

To make a new dictionary between two uncommon languages, it is often necessary to transform published dictionaries, one between the source and the international language, the other between the international and the target language. The problem in this process is to eliminate the irrelevant translations introduced by words with ambiguous meanings (Tanaka, 1994).

This can be thought of as **choosing the translations from several candidates without aligned corpus**. Note that adopting aligned corpus of insufficient size cause the same situation. We therefore propose a method to extract lexical translations using two corpora which are *not* aligned in the source and target language. Our method is proposed as the extension of the framework to solve the problem of choosing the translation according to the context. Thus, one of the merits of our research is that two problems, looking for the translation according to the global and local context, are handled within the same framework.

2 Assumption and Ambiguity Resolution

The source language is denoted as L_A and the target as L_B. Japanese and English have been adopted as L_A and L_B, respectively. Matrix A is defined with its (i,j)-th element as the value representing co-occurrence between two words a_i and a_j in L_A, with a similar definition for B. A and B are symmetric matrices. The number of words in L_A and L_B are denoted as N_A and N_B. The (i,j)-th element of matrix X is denoted as X_{ij}.

The cited Japanese examples are listed in the Appendix with their transliterations and first meanings. The cited English examples are written in *this font*.

2.1 Formalization

Translations of two co-occurring words in a source language also co-occur in the target language is assumed. For example, *doctor* and

*Author's current address: Department of Computer Science, Tokyo University of Agriculture and Technology. 2–24–16 Naka-machi, Koganei, Tokyo 184 JAPAN.

$$\begin{array}{ccc} & T & \\ a_u & \longrightarrow & b_k \\ A\updownarrow & & \Updownarrow T^tAT \text{ vs. B} \\ a_v & \longrightarrow & b_l \\ & T & \end{array}$$

Figure 1: Calculation of T^tAT

nurse co-occur in English and their translations 医者 and 看護婦 also co-occur in Japanese.

Rapp (1995) verified this assumption between English and German. He showed that two matrices A and B resemble each other, when a_i correspond to b_i for all i. Thus, the research had the additional assumption that English words and German words correspond one-to-one.

We introduce the translation matrix T from A to B because a word corresponds to several words rather than one. The (i,j)-th element of T is defined as the conditional probability $p(b_j|a_i)$, the translational probability of b_j given a_i. T forms a stochastic matrix, such that the sum of all elements in the same row is 1.0.

The co-occurrences A_{uv} in L_A can be translated into L_B using both $p(b_k|a_u)$ and $p(b_l|a_v)$:

$$\sum_{u,v} p(b_k|a_u)A_{uv}p(b_l|a_v) \quad (1)$$

Denoting for all B_{kl}, (1) can be rewritten in a simple matrix formulation as follows:

$$T^tAT \quad (2)$$

Note that the resulting matrix is also symmetric.

Returning to the example of *doctor* given in this section, its translation is 医者 but not 博士, because 看護婦, the translation of the co-occurring word *nurse*, co-occurs with 医者 but not with 博士. Thus, our assumption serves to resolve ambiguity.

This fact indicates that the translated co-occurring matrix T^tAT should resemble B (Figure 1). Defining $|X-Y|$ as a certain distance between matrices X and Y, ambiguity resolution is possible by simply obtaining T which minimizes the following formula:

$$F(T) = |T^tAT - B| \quad (3)$$

when A and B are known. Note that the above formulation assumes that the co-occurrence in L_A can be transformed congruently into L_B. Thus, T gives the pattern matching of two structures formed by co-occurrence relations (Section 4.2).

2.2 The Choice of Co-occurrence Measure and Matrix Distance

There are many alternatives to measure co-occurrence between two words x and y (Church, 1990; Dunning, 1993). Having $freq(x)$ as the count of x in the entire text, $freq(x,y)$ as the number of appearances of both x and y within a window of a fixed number of words, and N as the number of words in the text concerned, we adopt the following mutual information:

$$\frac{Nfreq(a_i,a_j)}{freq(a_i)freq(a_j)} \quad (4)$$

Rapp argues that $freq(a_i,a_j)^2/freq(a_i)freq(a_j)$ is although more sensitive than above. Formula (4), however, will be adopted due to its statistical property being already studied (Church, 1990).

Rapp normalized matrices A and B. We, however, do not normalize from the reason that the value by Formula (4) is already normalized by N[1].

Distance for matrices should also be considered. Rapp used the sum of absolute distance of the elements. Since our requirement is that the distance is easy to handle analytically to obtain T as in Section 4.1, the following definition was chosen:

$$|X-Y| = \sum_{i,j}(X_{ij}-Y_{ij})^2 \quad (5)$$

3 Local Ambiguity Resolution

Note that the elements with value 0.0 in a matrix are denoted by "–" in the following discussion.

3.1 Example of *doctor*

Suppose that *doctor* occurs in the local context "*The doctor nursed the patient.*" We want to disambiguate the meaning of *doctor* as the *medical doctor*, not *Ph.D.* As *doctor* co-occurs with *nurse* and *patient*, *nurse* with *doctor* and *patient* etc., the matrix A can be defined by Formula (4) as follows[2]:

	doctor	*nurse*	*patient*
doctor	–	3.0	3.0
nurse	3.0	–	3.0
patient	3.0	3.0	–

For T, only the ambiguity of *doctor* is concerned here for simplicity, not that of *nurse* or *patient*, giving T as follows:

	医者	看護する	患者	博士	大学
doctor	T_{11}	–	–	T_{41}	–
nurse	–	1.0	–	–	–
patient	–	–	1.0	–	–

Note that 大学 is a co-occurring word with 博士. Here we are interested in whether $T_{11} = 1.0$ (*doctor* — 医者) or $T_{41} = 1.0$ (*doctor* — 博士): the correct answer is clearly $T_{11} = 1.0$.

[1]When we renormalized A and B and applied the incremental calculation which will be indicated in Section 4, T empirically oscillated and did not converge, because N_A and N_B can differ drastically.

[2]The value 3.0 refers to N_A, which is calculated as $(N_A \times 1)/(1 \times 1) = N_A$. whereas 1 is the frequency of each occurrence. Here N_A is 3, the three words *doctor*, *nurse* and *patient*.

The quality of A is poor from a statistical point of view (Church, 1990). What is needed in the local ambiguity resolution is only the information of co-occurring words, and the co-occurrence values are not that important when forming A. Although there are other solutions for forming A, for example, to put all elements concerned simply to 1.0, this definition was used because the local and global problems can be handled within exactly the same framework.

B is obtained globally from the corpus in L_B. Suppose that B for the words in question is given for simplicity as follows:

	医者	看護する	患者	博士	大学
医者	–	10.0	50.0	–	–
看護する	10.0	2.0	8.0	–	–
患者	50.0	8.0	–	–	–
博士	–	–	–	3.0	15.0
大学	–	–	–	15.0	3.0

We experimentally put $T_{11} = 1.0$, so that *doctor* corresponds to 医者, and calculated T^tAT giving the following result with $F(T) = 5038$:

	医者	看護する	患者	博士	大学
医者	–	3.0	3.0	–	–
看護する	3.0	–	3.0	–	–
患者	3.0	3.0	–	–	–
博士	–	–	–	–	–
大学	–	–	–	–	–

Next, we put $T_{41} = 1.0$, so that *doctor* corresponded to 博士. T^tAT gave the following result with $F(T) = 5758$:

	医者	看護する	患者	博士	大学
医者	–	–	–	–	–
看護する	–	–	3.0	3.0	–
患者	–	3.0	–	3.0	–
博士	–	3.0	3.0	–	–
大学	–	–	–	–	–

These two results indicate that T with $T_{11} = 1.0$ (*doctor* — 医者) makes T^tAT and B closer than T with $T_{41} = 1.0$ (*doctor* — 博士). Therefore the translation of *doctor* is determined to be 医者.

The algorithm to choose the translation from several candidates reflecting the local context is summarized as follows:

1. Create a local A.
2. Make a T that assumes one candidate to be the translation. Calculate the distance $F(T)$ for each candidate.
3. Choose the T with the minimum $F(T)$.

3.2 Related Work

Dagan (1994) proposed a method to choose a translation according to the local context. The significance of this work is that the ambiguity is *not* solved within L_A, as was traditionally studied, but was solved in L_B, same as our standpoint. Word to be translated (a_u) and its relating word (a_v) concerning phrasal structure (for example objective for verb) were translated into L_B (b_i and b_j, respectively), using an electronic dictionary.

The co-occurring frequency within L_B was measured and $p(b_k, b_l|a_u, a_v)$ was estimated as follows:

$$p(b_k, b_l|a_u, a_v) = \frac{freq(b_k, b_l)}{\sum_{i,j} freq(b_i, b_j)} \quad (6)$$

Dagan chose b_k of the largest $p(b_k, b_l|a_u, a_v)$ as translation after statistically testing its reliability.

The difference with our method is that he estimated the translational probability between pairs (the word and its co-occurrence) whereas our framework reduces the translational probability of pairs into that of words. Thus, our method can be applied to obtain global translations, which will be explained in the following section.

4 Global Extraction of Translations

The extraction of global lexical translations is formulated using the same framework as ambiguity resolution in the local context. The difference is that A is formed globally from the corpus in L_A.

For local context, the number of possible translations is small enough that each case can be tested one after another to find the best T. Unfortunately, the same method cannot be applied to obtain global translations because the number of combinations of possible translations explodes. Hence, we propose a method to update T incrementally.

4.1 Steepest Descent Method

T is not a square matrix and the number of equations obtained by $T^tAT = B$ is not always equal to that of variables T_{ij}, so the equation may not be solved directly. We therefore try to obtain the best T by the Steepest Descent Method (SDM) to minimize the Formula (3). T is incrementally updated from T_n to T_{n+1} by:

$$T_{n+1} = T_n + dT \quad (7)$$

where dT can be calculated with ds being a certain small length as:

$$dT_{ij} = -\frac{\partial F}{\partial T_{ij}} ds \quad (8)$$

The result can be represented as follows:

$$dT = -4AT(T^tAT - B)ds \quad (9)$$

The constraint for T that the sum of the same row must be 1.0 can be reflected on the calculation using Lagrange's method of indeterminate coefficients.

4.2 Characteristics of Our Method

If words are regarded as nodes, relations such as co-occurrences and translations as branches, then matrices A, B and T represent graphs.

Suppose that A and B are exactly the same graph as in Figure 2. The representation matrices are also indicated in the figure.

The best T is obviously as follows,

$$T = \begin{pmatrix} – & – & – & 1.0 \\ – & – & 1.0 & – \\ – & 1.0 & – & – \\ 1.0 & – & – & – \end{pmatrix}$$

This means that a_1, a_2, a_3, a_4 correspond to b_4, b_3, b_2, b_1 respectively. It also indicates that a_1

$$A = \begin{pmatrix} - & p & q & - \\ p & - & r & s \\ q & r & - & - \\ - & s & - & - \end{pmatrix} \quad B = \begin{pmatrix} - & - & s & - \\ - & - & r & q \\ s & r & - & p \\ - & q & p & - \end{pmatrix}$$

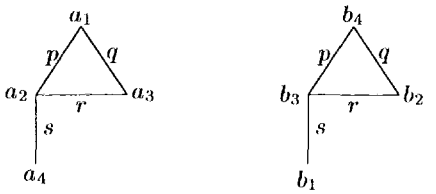

Figure 2: Graphs of Matrices A and B

$$B = \frac{1}{4}\begin{pmatrix} A & A \\ A & A \end{pmatrix}$$

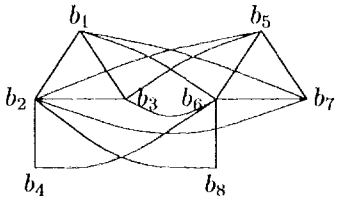

Figure 3: Another Graph of Matrix B

does not correspond to b_3, b_2, or b_1, which is exactly the disambiguation. In terms of linear algebra, the calculation $T^t AT$ is so-called a "congruent transformation." T provides the pattern matching of the two graphs given by A and B.

Next, suppose that A is defined as above and B is written in a block matrix as shown in Figure 3, containing the same graphs as A. T will clearly be $T = 1/2(E\ E)$ with E being a unit matrix of size 4. The point is that our algorithm has a limit for ambiguity resolution especially when there are several resembling graphs interconnected, that is, the ambiguity of a_1 cannot be resolved between b_1 and b_5.

On the other hand, as shown in (Brown, 1993), methods using aligned corpus does not have this limit. Starting his method with every English word corresponding to all French words, only several French words remain as translations in the result. This difference shows our weak point compared with Brown's.

Our method, assuming that two graphs can be linearly transformed, only tries to make a match between two graphs in L_A and L_B without aligned corpus, so some hints for obtaining the correct correspondences, some compensations for the lack of aligned corpus, are needed. For example, when the value of (i, j)-th element is zero in T_0, the value of the same element can be kept at zero during the SDM.

4.3 Related Work

Some research using aligned corpus point out problems with corpus size and noise, which leads to insufficient accuracy in translations.

Fung (1995) asserts that translation of words or phrases might not exist even in the aligned corpus. She extracted noun translations from noisy aligned corpus. First, a number of obvious translations were statistically extracted, then the uncertain translations were found using the co-occurrence with the obvious ones.

Utsuro (1994) claimed that there is a need to extract lexical translations even from an aligned corpus of a small size and proposed to use an electronic dictionary as an aid. First, a certain number of candidates are found. If a candidate in L_B co-occurs with another found in the electronic dictionary, its probability of being the translation is adjusted to be higher.

The common idea in the two approaches, the use of lexical co-occurrence within L_B, was also introduced by Dagan (1994).

Table 1: Local Ambiguity Resolution Power

POS	*correct*	*wrong*	*unresolved*
noun	93	20	27
verb	21	5	6
adjective	9	1	12
adverb	1	1	4
total	124	27	49

5 Experiments

Two experiments, local and global, were performed by choosing the Japanese translations for English words. The corpora adopted are the 30M Wall Street Journal and 33M political and economic articles of Asahi Newspaper.

These were morphologically analyzed[3] to extract nouns, verbs, adjectives and adverbs in canonical forms. Co-occurrences were counted using an 11 word window size. A and B were created as was depicted in Section 2.1. Elements under the certain thresholds were set at 0.0. The initial bilingual dictionary used was Edict (Breen, 1995), a word-to-word public dictionary.

5.1 Local Ambiguity Resolution

We randomly extracted 11 successive words from corpus. If the 6th center word was ambiguous satisfying the following three conditions, the method explained in Section 3.1 was applied for disambiguation: its translations could be subjectively judged according to the context; the translations exist in Edict; Edict contains candidates other than the translation.

The calculation choice was selected as the one which exhibited the minimum $F(T)$. If all the scores were the same, it was judged *unresolved*. When our subjectively judged translations contained the calculation choice, it was *correct*, otherwise *wrong*. The experiment was performed until the ambiguity was resolved for 200 different words.

Table 1 shows the results. The *applicability*, the rate of words which were not *unresolved*, was

[3]PC-KIMMO and JUMAN were used.

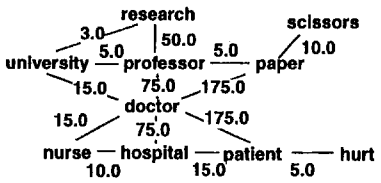

Figure 4: A Graph of *doctor*

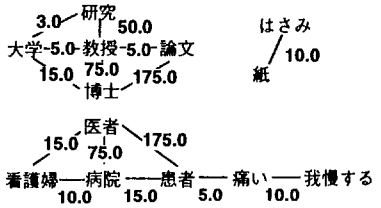

Figure 5: A Graph of 医者 and 博士

75.5% ((124+27)/200). The *correctness* (precision), the rate of the *correct* candidates among the words not *unresolved*, was 82.1% (124/(124+27)).

The general trends found are as follows:

- Translations reflect the trends in the corpus. For example, for *doctor*, 医師 was calculated to be the best choice. Although 医者 was also a candidate meaning *medical doctor*, it was dropped, because 医者 is a rather uncommon usage in the corpus.
- Most words with two obviously different meanings were calculated to obtain the *correct* result.

The *applicability* depends on the window size, such that the window should be large enough to focus the meaning of the word in question. The smaller the size is, the lower the rate should be. However, even if the window is made wider, the rate should eventually reach a certain limit.

5.2 Global Extraction of Translations

Example of *doctor*

Figure 4 shows a small graph concerning *doctor*. The values attached to branches represent co-occurrences. Figure 5 shows the corresponding graph in Japanese. We initially defined A and B from these graphs, and T_0 as each English word corresponding one-to-one to the Japanese word (with a value 1.0), except that three ambiguous words have the following correspondences:

doctor	→	医師 (0.333), 博士 (0.333), 教授 (0.334)
patient	→	我慢する (0.5), 患者 (0.5)
paper	→	論文 (0.5), 紙 (0.5)

SDM was applied to T_0 and its convergence was judged with the first 5 digits of $F(T)$. This needed 3400 iterations for convergence. The result T_{3400} is as follows:

doctor	→	医師 (0.502), 博士 (0.498), 教授 (0.0)
patient	→	我慢する (0.0), 患者 (1.0)
paper	→	論文 (0.989), 紙 (0.011)

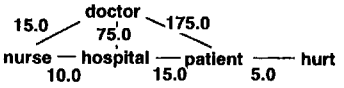

Figure 6: A Graph of *medical doctor*

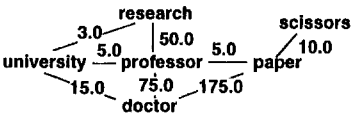

Figure 7: A Graph of *Ph.D.*

The wrong translation *doctor*—教授 was dropped.

Next, we removed from Figure 4 the portion of the graph which corresponds to the meaning of *Ph.D.* (Figure 6) so that the context was restricted to *medical doctor*. This time the result was:

doctor	→	医師 (1.0), 博士 (0.0), 教授 (0.0)
patient	→	我慢する (0.0), 患者 (1.0)

Then we removed from Figure 4 the portion of the graph which corresponded to the meaning of *medical doctor* (Figure 7) so that the context was restricted to *Ph.D*, giving the result:

doctor	→	医師 (0.0), 博士 (1.0), 教授 (0.0)
paper	→	論文 (0.996), 紙 (0.004)

These three small experiments show that the translation for *doctor* reflects the context represented by the source graph in L_A.

Minor Analysis of 378 words

The best experiment is to calculate T for entire dictionary and measure how much the obtained translations reflect the corpus context, but this is difficult both from calculation time and judgment of context reflection. Hence we intentionally added to Edict the irrelevant translations to see if they drop out by our method.

The irrelevant translations were chosen randomly so that they become the same number as those which existed originally in Edict. This was performed for entire English words in Edict. A was formed so that all the words involved are reachable within 2 co-occurrence branch distances from the test word. B is created by all translations of words involved in A. The test words applied SDM was selected by the following conditions: a test word has more than one candidate (ambiguous words) in Edict; its all co-occurrence values are greater than a certain threshold.

If the candidates are separated into the following three categories through calculation: those which gain value, decrease value, and those whose values do not change, then we define the word in question as *applicable*. The following rates were calculated for CDIW (correctly dropped irrelevant words, the irrelevant words added as a noise and dropped correctly by the method) for each *applicable* test words:

Table 2: Dropped Irrelevant Translations

threshold	*applicability*	*correctness*	*coverage*
50.0	68.3%	84.7%	35.2%
30.0	84.7%	84.6%	41.9%

- The fraction between the number of CDIW and dropped words. (*correctness*, recall)
- The fraction between the number of CDIW and irrelevant words. (*coverage*)

The results are listed in Table 2.

The *applicability* and *coverage* depend on the threshold: the lower the threshold is, the higher the two rates increase because more co-occurrence information is obtained. The threshold is a trade-off with calculation time.

About 15% (100−84.6) incorrectly dropped ones were original translations contained in Edict. These did not match the context, similar to the case of (*doctor*—医者) shown in Section 5.1.

6 Conclusions

Lexical translations were extracted from non-aligned corpora. The assumption that "translations of two co-occurring words in a source language also co-occur in the target language" was introduced and represented in the stochastic matrix formulation. The translation matrix provides the co-occurring information translated from the source into the target. This translated co-occurring information should resemble that in the target when the ambiguity of translational relation is resolved. This condition was used to obtain the best translation matrix.

The proposed framework, aimed at ambiguity resolution, serves to globally obtain lexical translations using non-aligned corpora just as to choose a translation according to the local context. The algorithms for obtaining the best translation matrix were shown based on the Steepest Descent Method, an algorithm well known in the field of non-linear programming.

Two experiments were performed to examine the power of local ambiguity resolution and dictionary refinement. The former showed a precision of 82.1% with applicability of 75.5%. In the latter, irrelevant translations were intentionally added to the dictionary to examine whether the relevant ones will be chosen. It was found that 84.7% of the dropped words were indeed irrelevant ones.

An important future task is to decrease the computational complexity. The method is applicable to matrix calculation with the size of an entire dictionary, but this is unrealistic at this stage. We must also increase the rate of ambiguity resolution. The corpus is regarded as non-structured data in this paper, the ambiguity might be resolved more effectively by introducing a phrasal structure.

Acknowledgment

We thank Dr. Koiti Hasida for useful discussion. Our experiments are supported by Dr. Kyoji Umemura's corpus data. We express our gratitudes to Mr. Breen for providing his Edict for our experiments.

References

James W. Breen, (1995). Edict, Freeware Japanese / English Dictionary.

Peter F. Brown et al. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, vol. 19 (2), pp. 263–311.

Kenneth W. Church and Patrick Hanks (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, vol. 16 (1), pp. 22–29.

Ido Dagan and Alon Itai (1994). Word Sense Disambiguation Using a Second Language Monolingual Corpus, *Computational Linguistics*, vol. 20 (4), pp. 563–596.

Ted Dunning (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics*, vol. 19 (1), pp. 61–74.

Pascale Fung (1995). A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. *Proceedings of ACL '95*, pp. 236–243.

Reinhard Rapp (1995). Identifying Word Translations in Non-Parallel Texts. *Proceedings of ACL '95*, pp. 321–322.

Kumiko Tanaka and Violaine Prince (1995). Amelioration Automatique Incrémentale de Dictionnaires Bilingues Utilisant un Corpus Monolingue. *Conference Internationale d'AUPELF '95.*

Kumiko Tanaka and Kyoji Umemura (1994). Construction of a Bilingual Dictionary Intermediated by A Third Language. *Proceedings of the International Conference for Computational Linguistics '94*, pp. 293–393.

Takehito Utsuro et al. (1994). Bilingual Text Matching using Bilingual Dictionary and Statistics. *Proceedings of the International Conference for Computational Linguistics '94*, pp. 1076–1082.

Appendix

Japanese	Transliteration	First meaning
医者	isha	medical doctor
医師	ishi	medical doctor
博士	hakase	Ph.D.
看護婦	kangohu	nurse
看護する	kangosuru	to nurse
患者	kanja	patient
痛い	itai	hurt
大学	daigaku	university
論文	ronbun	paper as articles
教授	kyouju	professor
我慢する	gamansuru	be patient
紙	kami	paper to write on
はさみ	hasami	scissors