Q1

We consider POS tagging of an English sentence "break the green wall." We rank solutions of POS tagging with Hidden Markov Model (HMM), and use Viterbi algorithm to obtain the most likely solution. Fig. 1 is a word dictionary, while Fig. 2 and Fig. 3 are parameters of HMM. In the table of Fig. 3, raws stand for C_i and columns C_j (for example, $P(N|D) = 0.8$). ϕ is a beginning of sentence.

Word	POS
break	N, V
green	A, N
the	D
wall	N, V

Fig. 1 Word Dic.

$$
\begin{aligned}
P(green|A) &= 3.0 \times 10^{-6} \\
P(the|D) &= 0.5 \\
P(break|N) &= 2.0 \times 10^{-6} \\
P(green|N) &= 1.0 \times 10^{-6} \\
P(wall|N) &= 1.0 \times 10^{-5} \\
P(break|V) &= 5.0 \times 10^{-5} \\
P(wall|V) &= 2.0 \times 10^{-6}
\end{aligned}
$$

Fig. 2 $P(w_i|C_i)$

	A	D	N	V
ϕ	0.1	0.6	0.2	0.1
A	0.1	ⓐ	0.7	0.1
D	0.2	0	0.8	0
N	0.1	0.3	0.1	0.5
V	0.1	0.6	0.2	0.1

Fig. 3 $P(C_j|C_i)$

1. Answer the probability $P(D|A)$. (In other words, fill ⓐ in Fig.3.)
2. Draw the word lattice. Show also the probabilities of nodes and links. Furthermore, show local best probabilities of nodes obtained by Viterbi algorithm, and links of the path corresponding to the local best probabilities.
3. Answer the most likely solution and its probability.

Q2

Suppose that an optical character reader (OCR) accepts a handwritten word, and outputs W1 and W2 below as candidates of solutions.

W1. power W2. qovei

1. For each W1 and W2, answer the formula of generation probability of the word when it is approximated as bi-gram (2-gram) model of characters.
2. The tabel in right is a part of 2-gram statistics obtained from a corpus. At the cell of i-th row and j-th column in the table, there is a number of times such that j-th character appears just after i-th character. ϕ denotes the beginning of a sentence. Calculate the generation probability of W1 and W2 following the formula you gave in the previous question 1. Then choose the correct solution according to the calculated generation probabilities.

	⋯	e	i	o	p	q	r	v	w	⋯	(Total)
ϕ	⋯	60	50	30	80	25	400	50	40	⋯	10000
⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮
e	⋯	10	20	40	10	4	500	20	30	⋯	20000
i	⋯	100	5	60	30	2	40	50	10	⋯	10000
o	⋯	20	20	5	50	1	100	60	45	⋯	15000
p	⋯	250	200	300	5	0	80	20	30	⋯	6000
q	⋯	10	1	9	0	0	0	0	0	⋯	30
r	⋯	200	300	200	2	0	5	15	20	⋯	8000
v	⋯	40	50	20	3	0	10	0	0	⋯	400
w	⋯	42	40	50	5	0	15	0	0	⋯	600
⋮		⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		⋮

3. **Group A** shows three tasks based on statistical analysis of languages.

Group A

A1. Improvement of speech recognition A2. Word segmentation without a dictionary A3. Similarity test between texts

(a) From Group A, choose all tasks for which n-gram model is used.

(b) From Group A, choose all tasks for which χ^2 test is used.

Q3

For each task in (a) $\sim$ (e) below, choose the most appropriate corpus used for it from **Group B**. If there is no appropriate corpus, just answer 'nothing'.

Group B

B1. plain text, B2. part-of-speech tagged corpus, B3. tree annotated corpus, B4. sense tagged corpus, B5. parallel corpus

(a) Estimation of parameters of probabilistic context free grammar.

(b) Estimation of parameters of translation model $P(S|T)$ in statistical machine translation.

(c) Estimation of parameters of language model $P(T)$ in statistical machine translation.

(d) Evaluation of a parser (syntactic analysis system).

(e) Example based case analysis.

Q4

1. There are 4 major methods for machine translation (MT).

   (a) "Interlingua Method" is suitable for multilingual MT system. Explain the reason why.

   (b) "Transfer Method" consists of three modules: **Transfer**, **Analysys** and **Generation**. Put them in the order of execution. (Answer the module executed first, second and third.)

   (c) Choose all knowledge required for "Example based MT" from **group C**.

   group C

C1. grammar C2. case frame dictionary C3. thesaurus C4. translation memory C5. syntactically annotated corpus

   (d) "Statistical machine translation" is based on the probabilistic model $P(S|T)P(T)$. In what point of view does $P(S|T)$ or $P(T)$ evaluate quality of a translated sentence T? Answer for both $P(S|T)$ and $P(T)$.

2. **Group E** is a list of fundamental techniques in a dialog system. Which techniques does the sentence (a) $\sim$ (f) explain? Choose all techniques from **group E** for each. If the sentence explains none of them, just answer 'nothing'.

   Group E

E1. understanding of user's utterance E2. planning E3. plan recognition E4. discourse structure analysis

   (a) It is a process to understand an user's intention.

   (b) It is a process to decide a content of a system's utterance.

   (c) It is a process to detect change of topics in a dialog.

   (d) Prosody (pitch, intonation, volume etc.) is considered.

   (e) Morphological, syntactic and semantic analysis are used.

   (f) It utilizes causal rules prepared in advance.

Q5

We consider text retrieval to know what kinds of bear are in 'red list' (list of threatened species). Fig. 4 shows document collection used for text retrieval and index terms in each document. D_i stands for a document, while a number in ( ) stands for number of times that the index term appears in that document.

D_1	red(2), list(1), polar(1), bear(1)
D_2	brown(1), bear(2)
D_3	brown(1), polar(1), bear(2)
D_4	red(1), list(1), panda(2)
D_5	red(3), teddy(1), bear(1)

Fig. 4 Document Collection

	D_1	D_2	D_3	D_4	D_5
①	0	0	0	0	1
②	0	0	0	1	0
③	1	1	1	0	1
?	0	1	1	0	0
?	1	0	1	0	0
?	1	0	0	1	0
?	1	0	0	1	1

Table 1

$$\vec{q} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \begin{matrix} (bear) \\ (brown) \\ (list) \\ (panda) \\ (polar) \\ (red) \\ (teddy) \end{matrix}$$

Fig. 5

1. Table 1 is the inverted index constructed from the document collection in Fig. 4. Fill ① ∼ ③ in Table 1.

2. We use Vector Space Model for text retrieval. A query vector $\vec{q}$ is defined as Fig. 5. Each dimension in the vector corresponds to an index term in parentheses ( ). Show document vectors $\vec{d_1} \sim \vec{d_5}$ for each document. Weights of index terms should be determined as TF·IDF. Use approximations in the table below.

x	1	1.25	1.33	1.5	1.66	2	2.5	3	4	5
$\log x$	0	0.22	0.29	0.41	0.51	0.69	0.92	1.1	1.4	1.6

3. We define $sim(\vec{q}, \vec{d_i})$, the similarity between $\vec{q}$ and $\vec{d_i}$, as Equation (1). Show the similarities $sim(\vec{q}, \vec{d_i})$ for documents $D_1 \sim D_5$. Answer the most similar document, too.

$$sim(\vec{q}, \vec{d_i}) = \vec{q} \cdot \vec{d_i} \qquad \text{(inner product of vectors)} \tag{1}$$

4. Query Expansion is a procedure to automatically add index terms which are related to w_q, where w_q is an index term whose weight in the query vector $\vec{q}$ is 1. Now we try query expansion using the thesaurus shown in Fig.6.

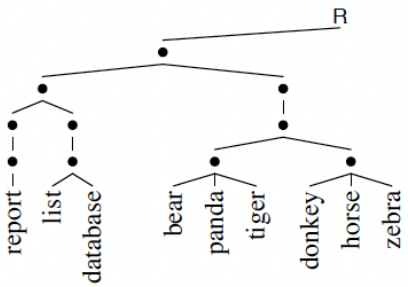


Fig. 6 Thesaurus

   (a) We define the similarity between two words $s(w_1, w_2)$ as Equation (2). Calculate the similarity $s(\text{bear}, \text{tiger})$, $s(\text{bear}, \text{zebra})$, $s(\text{list}, \text{database})$ and $s(\text{list}, \text{report})$.

$$s(w_1, w_2) = (\text{depth of the common superordinate node on the thesaurus}) \tag{2}$$

   (b) As query expansion, the words w_{ex} which fulfill all following requirements are added to the query.

   - $s(w_{ex}, w_q) \geq 4$. (i.e, w_{ex} is a word similar to w_q)
   - w_{ex} should appear once or more in the document collection in Fig.4.

   Answer a query vector $\vec{q'}$ after query expansion. Note that the weight of w_q should be 1, while the weight of the newly added index terms by query expansion (w_{ex}) should be 0.5.

5. 'Precision' and 'recall' are major evaluation criteria for information retrieval. Show the definition of them.

Q6

1. Fig.7 shows word dictionaries. In the column of 'key' in Fig.7 **F1**, Japanese words are put in alphabetical order.
   (a) Which of two dictionaries can be used when we search for a word by 'binary search' , **F1** or **F2**? If neither can, just answer 'nothing'.
   (b) Which of two dictionaries can be used when we search for a word by 'hashing' , **F1** or **F2**? If neither can, just answer 'nothing'.
   (c) What is the task of natural language processing for which a dictionary like Fig.7 **F2** (called TRIE) is especially useful?
2. The graph in Fig. 8 represents a certain law. Answer the name of that law. Furthermore, explain what is r in horizontal axis and f_r in vertical axis.
3. Fig. 9 illustrates the flowchart of a question answering system. Answer name of module of (i) $\sim$ (iii).
4. Answer the following questions about information extraction. Just brief explanation in one or two lines is fine.
   (a) What is 'frame' in information extraction.
   (b) In information extraction, how to extract desired information from a document?

F1

key	content
あける	V
あま	N
あまい	A
い	N
いす	N
いた	N
いたい	A, V
らっぱ	N

F2

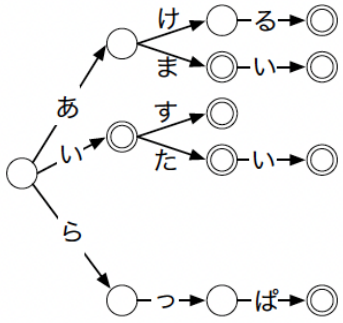


Fig. 7 Word Dictionary

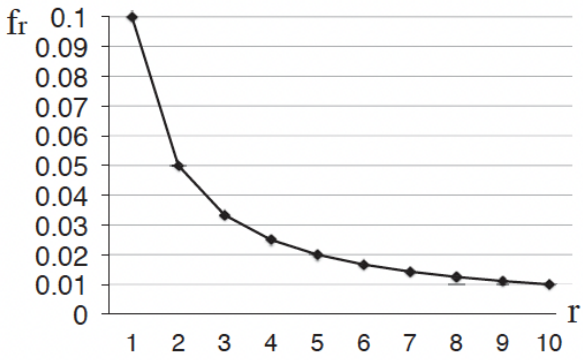


Fig. 8

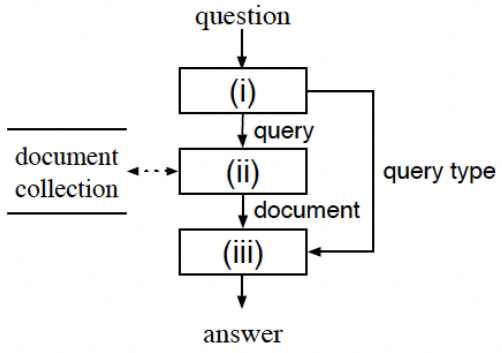


Fig. 9

Solution

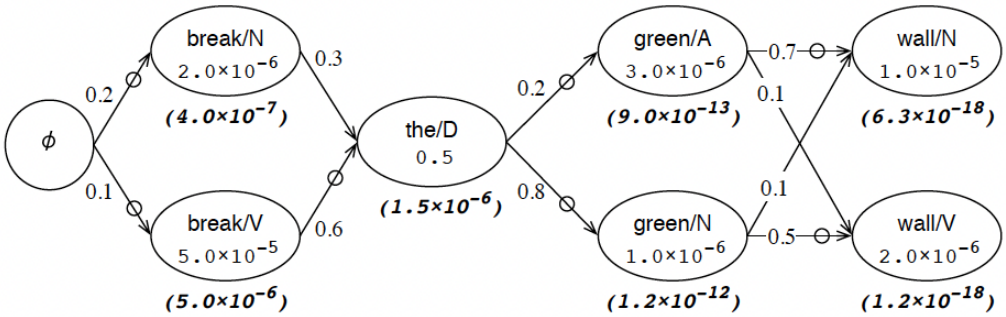
Q1

1. (2)

   0.1

   Because $\sum_{p\in\{A,D,N,V\}} P(p|A) = 0.1 + Ⓐ + 0.1 + 0.7 = 1.$

2. (12)



   (FYI)

   - NODE the/D

     $(4.0 \cdot 10^{-7}) \times (0.3) \times (0.5) = 6.0 \cdot 10^{-8}$

     $(5.0 \cdot 10^{-6}) \times (0.6) \times (0.5) = 1.5 \cdot 10^{-6}$ *

   - NODE wall/N

     $(9.0 \cdot 10^{-13}) \times (0.7) \times (1.0 \cdot 10^{-5}) = 6.3 \cdot 10^{-18}$ *

     $(1.2 \cdot 10^{-12}) \times (0.1) \times (1.0 \cdot 10^{-5}) = 1.2 \cdot 10^{-18}$

   - NODE wall/V

     $(9.0 \cdot 10^{-13}) \times (0.1) \times (2.0 \cdot 10^{-6}) = 1.8 \cdot 10^{-19}$

     $(1.2 \cdot 10^{-12}) \times (0.5) \times (2.0 \cdot 10^{-6}) = 1.2 \cdot 10^{-18}$ *

3. (2)

   break/V the/D green/A wall/N

   6.3×10^{-18}

Q2

1. (4)

   W1: P(p|ϕ) P(o|p) P(w|o) P(e|w) P(r|e)

   W2: P(q|ϕ) P(o|q) P(v|o) P(e|v) P(i|e)

2. (8)

   W1: $\frac{80}{10000} \times \frac{300}{6000} \times \frac{45}{15000} \times \frac{42}{600} \times \frac{500}{20000} = 2.1 \cdot 10^{-9}$

   W2: $\frac{25}{10000} \times \frac{9}{30} \times \frac{60}{15000} \times \frac{40}{400} \times \frac{20}{20000} = 3.0 \cdot 10^{-10}$

   W1 is correct. / W1 が正しい

3. (3)

   (a) A1, A2

   (b) A3

Q3

(5)

(a) B3　(b) B5　(c) B1　(d) B3　(e) nothing / なし

Q4

1. (a) (4)

   It is required to construct a module to convert between a language and interlingua. It is not required to construct translation module for all pairs of languages.

   言語と中間言語を変換するモジュールを作るだけでよく、全ての言語の組について翻訳モジュールを個別に作る必要がないため。

   (b) (2)

   Analysis → Transfer → Generation

   解析 → 変換 → 生成

   (c) (2)

   C3, C4

   (d) (4)

   $P(S|T)$: faithfullness (how likely T keeps the meaning of S)
   $P(T)$: fluency (how natural T is)

   $P(S|T)$: 正確性 (T がどれだけ S の意味を保持しているか)
   $P(T)$: 流暢さ (T がどれだけ自然な文か)

2. (6)

   (a) E3　(b) E2　(c) E4　(d) E4　(e) E1　(f) E2, E3

Q5

1. (3)

   ① teddy　② panda　③ bear

2. (9)

$$\vec{d_1} = \begin{pmatrix} 1 \times \log \frac{5}{4} \\ 0 \\ 1 \times \log \frac{5}{2} \\ 0 \\ 1 \times \log \frac{5}{2} \\ 2 \times \log \frac{5}{3} \\ 0 \end{pmatrix} = \begin{pmatrix} 0.22 \\ 0 \\ 0.92 \\ 0 \\ 0.92 \\ 1.02 \\ 0 \end{pmatrix} \quad \vec{d_2} = \begin{pmatrix} 2 \times \log \frac{5}{4} \\ 1 \times \log \frac{5}{2} \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.44 \\ 0.92 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

$$\vec{d_3} = \begin{pmatrix} 2 \times \log \frac{5}{4} \\ 1 \times \log \frac{5}{2} \\ 0 \\ 0 \\ 1 \times \log \frac{5}{2} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0.44 \\ 0.92 \\ 0 \\ 0 \\ 0.92 \\ 0 \\ 0 \end{pmatrix} \quad \vec{d_4} = \begin{pmatrix} 0 \\ 0 \\ 1 \times \log \frac{5}{2} \\ 2 \times \log \frac{5}{1} \\ 0 \\ 1 \times \log \frac{5}{3} \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0.92 \\ 3.2 \\ 0 \\ 0.51 \\ 0 \end{pmatrix}$$

$$\vec{d_5} = \begin{pmatrix} 1 \times \log \frac{5}{4} \\ 0 \\ 0 \\ 0 \\ 0 \\ 3 \times \log \frac{5}{3} \\ 1 \times \log \frac{5}{1} \end{pmatrix} = \begin{pmatrix} 0.22 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1.53 \\ 1.6 \end{pmatrix}$$

3. (4)

   $sim(\vec{q}, \vec{d_1}) = 0.22 + 0.92 + 1.02 = 2.16$
   $sim(\vec{q}, \vec{d_2}) = 0.44$
   $sim(\vec{q}, \vec{d_3}) = 0.44$
   $sim(\vec{q}, \vec{d_4}) = 0.92 + 0.51 = 1.43$
   $sim(\vec{q}, \vec{d_5}) = 0.22 + 1.53 = 1.75$

   The similarity to D_1 is maximum. / D_1 との類似度が最大。

4. (a) (4)

   $s(\text{bear}, \text{tiger}) = 4$
   $s(\text{bear}, \text{zebra}) = 3$
   $s(\text{list}, \text{database}) = 4$
   $s(\text{list}, \text{report}) = 2$

   (b) (3)

$$\vec{q'} = \begin{pmatrix} 1 \\ 0 \\ 1 \\ 0.5 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

5. (4)

$$\text{precision} = \frac{\text{number of relevant documents the system outputs}}{\text{number of documents the system outputs}}$$

$$\text{recall} = \frac{\text{number of relevant documents the system outputs}}{\text{number of relevant documents in overall collection}}$$

$$\text{精度} = \frac{\text{システムが出力した適合文書数}}{\text{システムが出力した文書数}}$$

$$\text{再現率} = \frac{\text{システムが出力した適合文書数}}{\text{文書集合における適合文書数}}$$

Q6

1. (a) (2)

   F1

   (b) (2)

   nothing / なし

   (c) (2)

   Morphological analysis of Japanese / 日本語の形態素解析

2. (4)

   Zipf's law / ジップの法則

   r is a rank (order) of frequency of words. f_r is a relative frequency of a word whose rank is r.

   r は単語のランク (出現頻度の順位)。f_r はランクが r の単語の相対出現頻度。

3. (5)

   (i) Question Analysis / 質問文解析
   (ii) Text Retrieval / テキスト検索
   (iii) Answer Extraction / 解答抽出

4. (a) (2)

   A table that defines the information to be extracted from a document.

   文書から抽出するべき情報を定義した表。

   (b) (2)

   Information is extracted by pattern matching.

   パターンマッチングによって抽出される。