PAPER Special Section on Natural Language Processing and its Applications

Robust Toponym Resolution Based on Surface Statistics

Tomohisa SANO[†a)], ***Nonmember***, **Shiho Hoshi NOBESAWA**[††], ***Member***, **Hiroyuki OKAMOTO**[†], **Hiroya SUSUKI**[†], **Masaki MATSUBARA**[†], ***Nonmembers***, ***and*** **Hiroaki SAITO**[†], ***Member***

SUMMARY Toponyms and other named entities are main issues in unknown word processing problem. Our purpose is to salvage unknown toponyms, not only for avoiding noises but also providing them information of area candidates to where they may belong. Most of previous toponym resolution methods were targeting disambiguation among area candidates, which is caused by the multiple existence of a toponym. These approaches were mostly based on gazetteers and contexts. When it comes to the documents which may contain toponyms worldwide, like newspaper articles, toponym resolution is not just an ambiguity resolution, but an area candidate selection from all the areas on Earth. Thus we propose an automatic toponym resolution method which enables to identify its area candidates based only on their surface statistics, in place of dictionary-lookup approaches. Our method combines two modules, area candidate reduction and area candidate examination which uses block-unit data, to obtain high accuracy without reducing recall rate. Our empirical result showed 85.54% precision rate, 91.92% recall rate and .89 F-measure value on average. This method is a flexible and robust approach for toponym resolution targeting unrestricted number of areas.

key words: *natural language processing, toponym resolution, area identification, statistical information*

1. Automatic Toponym Resolution

Many of previous researches on toponym resolution took dictionary-based approaches. However, when we face documents which may contain toponyms worldwide, there should be the problem of dictionary-unregistered toponyms. It is not realistic to gather all the toponyms in perfect detail in a gazetteer, or to combine gazetteers worldwide overcoming their difference in format. As a better and more realistic approach, we propose an automatic method to estimate area candidates for toponyms based only on their surface statistics. Our system concentrates on the extraction of area candidates, and an area identification task or an area disambiguation task may be required afterward.

2. Related Work

Identification of toponyms in documents generally requires the following two steps; toponym recognition (Sect. 2.1) and toponym resolution (Sect. 2.2).

Manuscript received March 20, 2009.
Manuscript revised July 8, 2009.

[†]The authors are with Keio University, Yokohama-shi, 223–8522 Japan.
[††]The author is with Tokyo City University, Tokyo, 158–8557 Japan.
a) E-mail: tomohisa@nak.ics.keio.ac.jp

DOI: 10.1587/transinf.E92.D.2313

2.1 Toponym Recognition

Named entity recognition, including toponym recognition, is a task to extract named entities embedded in documents. It has been one of the significant issues, concerning the unknown word problem.

Zhou et al. proposed an HMM-based chunk tagger based on the mutual information independence assumption for the named entity recognition problem [1]. Their method achieved F-measures of 96.6% for English named entity task.

There also is a maximum entropy tagger proposed by Curran et al. Their system extracted the required features with a language independent manner [2]. They resulted that their system works effectively not only for English but also for other languages.

2.2 Toponym Resolution

Toponyms sometimes appear in multiple areas. For example, *Portsmouth* is in U.K. and in U.S.A., and *Sparta* is in Greece and in U.S.A. Toponym resolution is a task to identify these multi-area toponyms, to estimate which *Portsmouth* the one in the target document is. Most of the toponym resolution approaches are based on a combination of gazetteers, context information, and information on the area.

Hauptmann and Olligschlaeger proposed rule-based methods [3], which are based on "one referent per discourse" assumption and spatial knowledge. This method achieved to resolve 269 toponyms correctly out of 357 toponyms.

Li et al. used a statistical approach to tag place names [4], which is based on a combination of pattern matching rules as well as discourse features based on co-occurring toponyms. The results were 96% accuracy on news articles and tourist guides.

Pouliquen et al. proposed a multilingual toponym resolution system based on a preference order [5]. In the disambiguation process, they used person name information, importance of the place, country information, geo-stop-word list and kilometric distance. It showed 0.77 of F-measure with the combination of the techniques for the toponym disambiguation.

Clough et al. proposed a toponym resolution system

based on multiple gazetteers [6]. They used pattern matching rules, ontologies and gazetteer preference orders to assign confidence scores for toponyms. They reported 89% accuracy focusing on the regions in the U.K., France, Germany and Switzerland.

Garbin and Mani used a statistical classifier for the toponym resolution [7]. They used various features, such as toponym classes, abbreviation information and capitalization of letters and collocated terms within a distance of three tokens from the target toponym. This automatically tagged data was used to train a machine learner, which disambiguated toponyms in a human-annotated news corpus at 78.5% accuracy.

Rauch et al. proposed a confidence-based toponym resolution system [8] using supportive or negative features such as the presence in gazetteer, linguistic context, etc.

These approaches require one or more fair gazetteers and heuristics methodologies for leveraging the context information. The quality of the systems depends on the knowledge base.

2.3 A Robust Approach to Toponym Resolution

For more robust processing, we should consider an approach which does not depend on toponym attributes, such as spatial data or population data.

Sano et al. proposed a statistical method for toponym resolution based on *n*-gram data [9]. Their system contained a set of binary classifiers, and the output was area candidates for an input toponym according to its surface information. They gained .93 of F-measure at maximum and .69 at minimum, though the method took a simple approach based only on letter-unit *n*-grams. The result showed that surface statistics can be effective in toponym resolution. But when a toponym does not contain sufficient information in its surface, the system failed to reduce the number of area candidates. This is because the system was designed for avoiding recall rate reduction, and the low precision rate was 55.09% at minimum.

Sano's system was robust as a candidate reducer, but it was not enough as a toponym identifier. Thus we propose a two-phased method, adding a TFIDF-based phase after Sano's SVM-based phase, for better performance as a toponym identifier by gaining more precision with a sufficient use of letter blocks [†], without decreasing recall rate.

3. Two-Phased Toponym Resolution Method

In this paper we propose an automatic estimation method of area candidates for toponyms, based on surface statistics, without looking up gazetteers. Our method takes a two-phased approach; (i) area candidate reduction phase based on letter-unit information and length information, and (ii) area candidate examination phase based on block-unit information. With these two phases, we achieved both high precision rate and high recall rate.

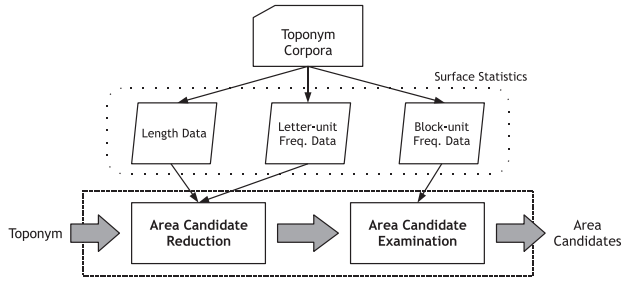

Fig. 1 Toponym resolution system.

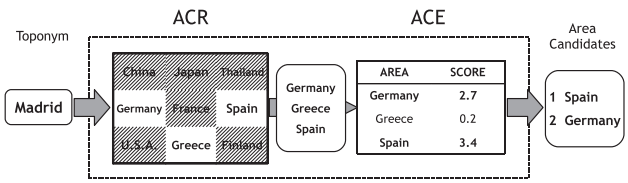

Fig. 2 Data flow.

3.1 Overview

The two phases which constitute our system are as follows: ACR, area candidate reduction phase which reduces area candidates with lower possibility, and ACE, area candidate examination phase which investigates the possibility of each remaining area candidate (Fig. 1). The top two rows in Fig. 1 show the prepared data. This system requires toponym corpora to extract the surface statistics. Each toponym corpus is a toponym list which contains toponyms for an area. As the preparation, we extract length data, letter-unit frequency data and block-unit frequency data from each toponym corpus. The bottom row in Fig. 1 shows the flow of our system. The first phase, ACR, outputs several area candidates out of all the areas prepared for the system. ACR reduces area candidates based on letter-unit frequency data and length data with high recall rate. And ACE ranks the area candidates in order of their possibility. In this ACE phase, it estimates the possibility of containing the input toponym for each area candidate by using TFIDF approach.

For example, when a toponym *Madrid* is given as an input, the system returns Spain and Germany as the output area candidates (Fig. 2). In this example, the set of possible area candidates contains nine areas. At ACR phase, three areas (Germany, Greece and Spain) out of nine remained as area candidates according to the letter-unit frequency data and the length data. At ACE phase, it ranks them according to their possibility, and then two area candidates with fair possibility are returned as system outputs. In this example, the possibility for Greece is too low to be an area candidate.

3.2 Area Candidate Reduction

Our first phase is an SVM-based system of Sano's ap-

[†]A block here is a kind of prefix and suffix. Detailed description is in Sect. 3.4.

Table 1 Toponym resolution with ACR.

area	F-measure	Precision (%)	Recall (%)
China	.85	75.64	97.94
Japan	.76	63.74	95.33
Thailand	.94	90.74	97.82
Greece	.70	56.69	91.22
Finland	.69	55.46	91.64
Germany	.79	66.30	98.39
France	.66	51.94	90.80
Spain	.72	59.64	91.29
U.S.A.	.80	68.86	94.31
TOTAL	.76	63.99	94.30

proach [9]. This system uses statistical information of the training toponym corpora as the features. As the support vector machine is a binary classifier, the system was made on one-versus-rest policy. That is, the system contains SVMs of the number of areas, each of which replies binary answers (positive or negative) for the area to an input toponym. The features they introduced were shown below. The total number of features was 30,833.

- letter-unit unigram, bigram and trigram
- number of letters and number of words in toponym
- number of words for each letter-unit length
- number of letters for each words

Each weight for a feature V is calculated with Expression (1) according to the n-gram frequency in an area A.

$$V_A(t, s) = P_A(s) \times N(t, s) \tag{1}$$

where s is an n-lettered string ($1 \leq n \leq 3$). $P_A(s)$ is the probability of s appearing in an area A, $N(t, s)$ is the frequency of a string s in a toponym t. The learning process of SVM is pre-processed for each toponym corpus. For each area, the toponyms of the area are positive instances and the others are negative instances. Table 1 shows the result of this ACR phase. With this phase, we can only obtain the list of area candidates. Recall rates are over 90% for all the nine areas, where the precision rates are much lower.

3.3 Area Candidate Examination

The second phase, ACE, calculates scores for each area candidate according to their possibility, and determines the output. Each area has its distinctive words. On the other hand, there also are common words among areas. In this phase the system calculates the possibility scores according to the appearance of these words.

For example, assume that there are four corpora for four areas, China, Finland, Japan and U.S.A., and now the system tries to estimate the area candidate for a toponym *Mt. Fuji San.* If a word *Fuji* is found only in Japanese training corpus but not in other areas' training corpora (Fig. 3), then the toponym may have strong possibility to be found in Japan. When all the area corpora were made of the same source, they may share the influence of the source language. For example, when the whole corpus set was made of a source written in English, we may need to consider the possibility that English words, such as *Mt.*, may be found in all the area corpora. In this case, the event containing a word *Mt.* should not be treated as an evidence of being a toponym of U.S.A. And also the possibility of being in Finland should not be changed, as there is no positive evidence to be in Finland and also there is no negative evidence.

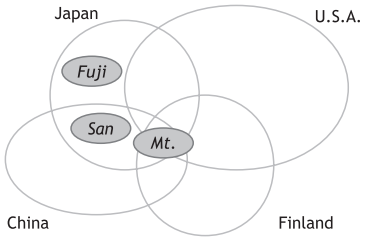

Fig. 3 Frequent words in toponym corpora.

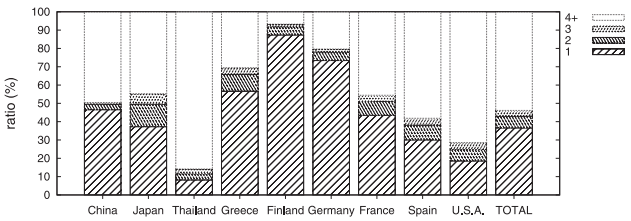

Fig. 4 Ratio of words by occurring numbers.

The possibility of a toponym being included in an area corpus, $E_A(t)$, can be defined as follows:

$$E_A(t) = \frac{1}{n} \sum_{k=1}^{n} e_A(w_k) \tag{2}$$

where t is a toponym, A is an area, w_k is the k-th word in the toponym, n is the number of words in the toponym and $e_A(w_k)$ represents the expectation value of a w_k being included in an area A. As the average number of words in a toponym is 1.94 (Sect. 4.1), we normalize the score with n to avoid the influence of the number of words (Expression (2)).

The expectation value $e_A(w_k)$ for an area A containing a word w_k is defined with TFIDF model as in

$$e_A(w_k) = TF_A(w_k) \log \frac{N}{DF(w_k)} \tag{3}$$

where $TF_A(w_k)$ is a frequency of the word w_k in an area A, and $DF(w_k)$ is the number of area corpora in the whole corpus set which contain the word w_k, and N is the number of area corpora. The system calculates the possibility score of each area candidate, and determines the area candidates with fair possibility.

3.4 Block-Unit Frequency Data

For the estimation of expectation value $e_A(w_k)$ in Expression (3) for the word w_k, we need to assume that the words contained in toponyms hold area-specific data sufficient enough for appropriate estimation. The words in toponyms, however, are mostly named entities, and the frequency of occurrence is not enough. Figure 4 shows the ratio of non-frequent words in toponyms. "1" in each bar shows the ratio

of words which occurred only once in each toponym corpus. "TOTAL" bar represents the ratio of words in the whole set of nine toponym corpora. More than 30% of the words in the toponym corpora appeared only once. These words do not exist in training corpora at open tests, thus the expectation values for these words are always zero. As most of the toponyms contain less than three words, even one unknown word may cause fatal error.

In this paper we propose a new unit, a block, in place of a word. We consider two types of blocks; prefix-type and suffix-type. For example, we may find a prefix-type block *yama* in Japanese toponym corpus as in *Yamagata*. And we may also find a suffix-type block *berg* in German toponym corpus as in *Heidelberg*. Though both *Yamagata* and *Heidelberg* are not frequent words, *yama* and *berg* should be frequent blocks.

Our method counts up these blocks instead of words, with which we can expect more frequent and area-specific surface data hidden in toponyms. Table 2 shows frequent five-lettered prefix-type blocks and suffix-type blocks in the toponym corpora. These are not actual prefixes or suffixes; they are presenting frequent n-grams at the beginnings and at the endings of words in toponyms. We can expect more frequency with blocks than words, which is the significant point in our surface statistics based approach. The longer block causes the data sparseness problem like word-unit frequency data have. Thus our method defines a block size as the maximum length of a block. When our method finds a word with shorter length than the block size, the method takes whole the word as a block. The prefix-type block and the suffix-type block are defined in Expression (4) and in Expression (5), respectively.

$$B_P(w_k, m) = c_1 \cdots c_i \quad (i = min(m, l)) \tag{4}$$

$$B_S(w_k, m) = c_j \cdots c_l \quad (j = max(l - m + 1, 1)) \tag{5}$$

where m is a block size, w_k is the word, c_i is the i-th letter of the word and l is the length of the word. Now we put a block-unit expectation value, in place of a word-unit expectation value $e_{A(w_k)}$ in Expression (3). Thus the possibility score, $E_A(t)$ in Expression (2), is here replaced with Expression (6).

$$E_A(t) = \frac{1}{n} \sum_{k=1}^{n} \frac{e_A(B_P(w_k, m)) + e_A(B_S(w_k, m))}{2} \tag{6}$$

Expression (6) enables the use of both the prefix-type blocks and the suffix-type blocks to represent the surface information of a toponym, instead of the direct use of words contained in a toponym.

Table 2 Frequent five-lettered blocks.

area	prefix-type block	suffix-type block
China	*hsien shang huang chang chuan chian shuik*	*huang hsien chang xiang hiang huiku cheng*
Japan	*shima machi misak nishi shimo higas shira*	*shima machi isaki betsu gashi shimo inami*
Thailand	*khlon ampho muang chang thung river luang*	*hlong muang mphoe thung luang river chang*
Greece	*ormos nisis vrach potam nisos nisid ayios*	*ormos nisis isida nisos aiika tamos orion*
Finland	*stora fjard sodra norra lilla svart vaste*	*jarvi niemi vaara saari lahti selka olmen*
Germany	*gross forst klein unter bahnh niede stein*	*ausen sdorf ingen ndorf forst nberg hnhof*
France	*saint rivie ruiss grand foret ville chate*	*saint ville viere sseau court foret ieres*
Spain	*arroy caser villa punta sierr barra corti*	*rroyo serio punta ierra ranco santa casas*
U.S.A.	*creek churc cemet schoo sprin branc mount*	*creek hurch etery chool ranch pring enter*

4. Empirical Result

4.1 Toponym Corpora

We used the toponym corpus set [†] shown in Table 3. #toponym, #word and #letter indicate the number of toponyms, the number of words and the number of letters included in each area corpus, respectively. We have no discrimination between lowercase and uppercase letters, and eliminate special letters like umlaut marks. This corpus set is made on English-language basis, thus these corpora may include English words in toponyms, such as *Elbe River* in Germany.

4.2 Effectiveness of Two-Phased Toponym Resolution with Block-Unit Frequency Data

Table 4 shows the empirical results of our proposing method

Table 3 Toponym corpus set.

area	#toponym	#word	#letter
China	10,000	15,975	98,773
Japan	10,000	16,960	102,311
Thailand	10,000	28,974	131,128
Greece	10,000	14,612	109,818
Finland	10,000	11,050	97,447
Germany	10,000	13,206	109,615
France	10,000	19,820	121,945
Spain	10,000	22,797	135,947
U.S.A.	10,000	25,622	165,509

Table 4 Toponym resolution results for each case.

case	F-measure	Precision (%)	Recall (%)
ACR	.76	63.99	94.30
ACEw	.67	92.96	52.77
ACEb	.87	90.51	83.84
ACR+ACEw	.82	72.87	93.37
ACR+ACEb	.89	85.54	91.92

[†]The toponym corpus set except U.S.A. corpus is created from the database of GEOnet Names Server (GNS) of National Geospatial-Intelligence Agency. The U.S.A. corpus is created from the data of Geographic Names Information System (GNIS) which is provided by United States Board on Geographic Names. GNS and GNIS contain approximately 5.5 million and 2 million names, respectively. We extracted 10,000 toponyms randomly from the database. In our experiment, we used only location name despite including many attributes such as longitude, latitude, etc.

Table 5 Empirical result of ACR+ACEb with possibility score ranking (%).

rank	China	Japan	Thailand	Greece	Finland	Germany	France	Spain	U.S.A.	TOTAL
1	86.2	88.7	96.1	79.4	80.2	86.2	73.8	80.4	92.0	84.8
2	1.7	1.2	1.0	2.4	1.8	1.6	6.3	3.3	1.4	2.3
3	.0	.0	.0	.1	.1	.1	.1	.1	.1	.1
4 – 9	.0	.0	.0	.0	.0	.0	.0	.0	.0	.0
unknown	10.0	5.4	.6	9.3	9.6	10.5	10.6	7.5	.9	7.1
negative	2.1	4.7	2.2	8.8	8.4	1.6	9.2	8.7	5.7	5.7

(ACR+ACEb), comparing to other four related methods (ACR, ACEw, ACEb, ACR+ACEw). ACR+ACEb stands for the combination of ACR and block-unit ACE, and ACR+ACEw stands for the combination of ACR and word-unit ACE.

We set 5 as the block size, for both the prefix-type blocks and the suffix-type blocks. For this experiment our system returns only top one area as the area candidate, according to the possibility. Based on the five-fold cross-validation on the corpora, we calculate the scores being in each area and compared the scores for each input toponym. Our method, ACR+ACEb, succeeded in obtaining high values both in precision (85.54%) and recall (91.92%), combining a high-recall area candidate reduction module and a high-precision area candidate examination module efficiently. Our block-unit examination approach was effective enough to cover the problem of data sparseness with word-unit examination approach.

Table 5 shows the result of ranking for all the toponyms in each toponym corpus (with an open test on cross-validation). Rank 1 in Table 5 shows the ratio of correct areas with top scores. "Unknown" in Table 5 indicates the ratio of toponyms which were decided as unknown toponyms (that is, no blocks in the toponyms found in the training corpora) at ACEb. There were 7.1% unknown toponyms in our system's output. "Negative" indicates the ratio of toponyms whose correct area were rejected as area candidates in ACR, and it contained 5.7% negatives. Table 5 shows that our system outputs correct area candidate with the top score for 84.8% of the toponyms.

5. Discussion

5.1 Word-Unit Frequency Data

Table 6 shows the example of the frequent words which appeared in most of the nine corpora. The number in the column DF shows the number of areas which contained the words. As shown in Table 6, most of the words with high document frequency are English words. This is because the toponym corpora were made on English gazetteer, and it contained English terms in toponyms, such as in *Biwako Canal*. Though included in Table 6, *de* and *la* gained more than 3% term frequency only in French toponym corpus and Spanish toponym corpus. These words can be regarded as area-specific words of France and Spain.

Table 7 shows the top 5 frequent words in each corpus. In Table 7, frequent words are mostly area-specific and English words appeared only in U.S.A. corpus. The distribution of frequent words is different among the area corpora. In the Spanish corpus and in the French corpus, the most frequent word is *de* and the second frequent word is *la*. These two are very frequent in these two corpora, and it is natural that these two area corpora show strong positive correlation.

Table 6 Examples of frequent words in whole corpus set.

DF	frequent words
9	*canal*
8	*river point*
7	*i bay island lake la de san*
6	*a e o station pass south cape channel north to west en*

Table 7 Examples of frequent words in each corpus.

area	frequent words
China	*shan hsien ho chen xian*
Japan	*yama mura saki gawa shima*
Thailand	*ban khlong huai khao nong*
Greece	*akra ormos nisis cape rema*
Finland	*iso stora sodra stor norra*
Germany	*berg bach bahnhof see wald*
France	*de la le saint les*
Spain	*de la del arroyo rio*
U.S.A.	*creek church cemetery school lake*

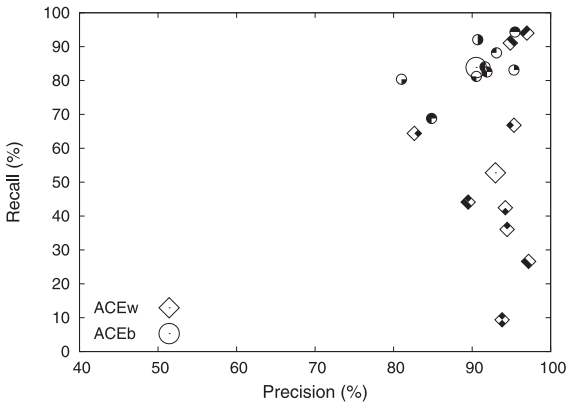

Fig. 5 Comparison between word-unit examination and block-unit examination.

5.2 Effectiveness of Block-Unit Frequency Data

Figure 5 shows the comparison between word-unit examination, ACEw and block-unit examination, ACEb at the second phase. The circles in Fig. 5 indicate block-unit results, and the diamonds indicate word-unit results, showing the overall results with big marks. Precision rates are

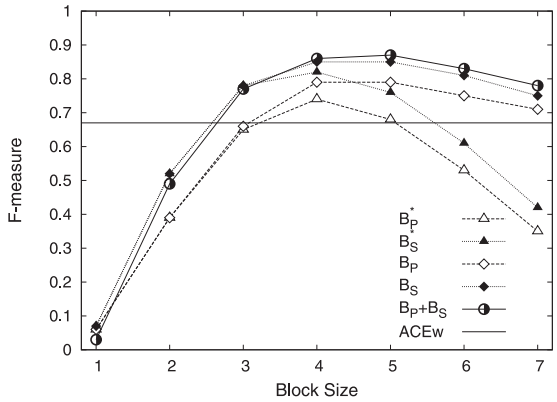

Fig. 6 Influence of block size and block usage type.

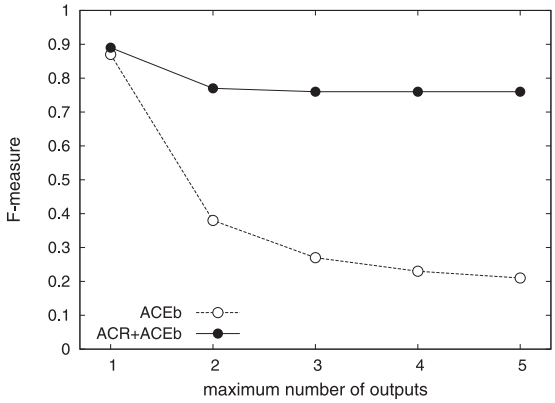

Fig. 7 F-measure values with multiple outputs.

good enough (around 90% in total) for both. There were difference in recall rates. The word-unit approach gained only 50% recall rate, while the block-unit approach obtained 80%. This difference is caused by the problem of data sparseness and unknown words in toponyms for the word-unit approach, which is described in Sect. 3.4. Thus we resulted that the use of the block-unit data instead of the word-unit data can be a key for the recall rate improvement without the precision rate reduction.

5.3 Consideration on Effective Use of Block-Unit Frequency Data

Here we describe the basic ideas on the definition of the blocks shown in Expression (4) and Expression (5) in Sect. 3.4. Fig. 6 shows the F-measure values according to the block sizes, block types (prefix-type and/or suffix-type), and the inclusion/exclusion of shorter words. Block size indicates the number of letters of a block. For example, block size 1 means that a block consists only the first/last letter of a word. When a word contains less letters than the block size, we call it a shorter word. For example, *de* is a shorter word for five-lettered blocks. Our proposing system includes these shorter words as blocks, for better performance by including short significant words in toponyms. We also show the difference of F-measure values by the inclusion and the exclusion of these shorter words in Fig. 6. The x-axis indicates the block size. And the y-axis stands for the F-measure value on ACE phase. There are five lines drawn in the graph, each stands for a different type of block usage described as follows:

B_P^*	Prefix-type blocks only, shorter words excluded
B_S^*	Suffix-type blocks only, shorter words excluded
B_P	Prefix-type blocks only, shorter words included
B_S	Suffix-type blocks only, shorter words included
$B_P + B_S$	Both prefix-type and suffix-type blocks, shorter words included

The horizontal line in the graph indicates the F-measure value .67 of the word-unit examination result. Inclusion of shorter words was also effective, especially at the experiments on longer blocks. Comparing the F-measure values of the prefix-type and the suffix-type, there is no significant difference between B_P^* vs. B_S^* and B_P vs. B_S. With this result we consider that both the prefix-type blocks and the suffix-type blocks have effectiveness at the same level for the scoring. The combination of the two, $B_P + B_S$, showed the best score for most of the block sizes. Thus we defined B_P and B_S as in Expression (4) and Expression (5), using both the prefix-type blocks and the suffix-type blocks, considering shorter words. The best block size should be 5, according to Fig. 6.

5.4 Number of Output Area Candidates

ACE is not designed to provide only one topmost-scored area candidate, thus we can consider multiple outputs according to the score ranking for better recall rates. To explain the influence of multiple outputs with ACE, we show the F-measure values for varied number of outputs in Fig. 7. The line with black dots shows the values with ACR+ACEb, and the line with white dots shows the values with ACEb. The x-axis indicates the maximum number of outputs at ACEb, and the y-axis indicates the F-measure values. For both approaches, F-measure values showed the best at one output and the F-measure values are almost the same at that point. The F-measure values show difference with bigger output numbers, which is caused by the decreasing of precision values with ACEb. As shown in Table 5, over 80% of toponyms come to Top 1 and it means that increasing the number of output area candidates causes decreasing the precision rather than increasing the recall. Thus we decided that our system returned only one candidate for the aspect of the precision rate.

5.5 Capability of Our Toponym Resolution Method

Table 8 shows the detailed result of our toponym resolution system ACR+ACEb. The result shows 85.54% precision rate, 91.92% recall rate and .89 of F-measure in total. The worst in F-measure was .81 with French toponym lists, and the best was .98 with Thai toponym lists. We consider that .81 of F-measure at the worst is successful enough with an

Table 8 Empirical result of ACR+ACEb.

area	F-measure	Precision (%)	Recall (%)
China	.92	88.61	96.23
Japan	.90	86.74	94.09
Thailand	.98	98.39	96.78
Greece	.86	82.79	88.66
Finland	.84	79.29	89.73
Germany	.90	83.89	96.68
France	.81	77.16	84.41
Spain	.85	82.86	87.86
U.S.A.	.92	92.12	92.85
TOTAL	.89	85.54	91.92

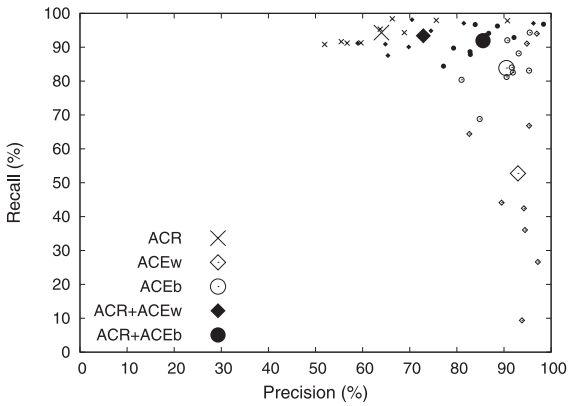

Fig. 8 Toponym resolution results for each area.

open test.

Figure 8 shows all the empirical results for the tested areas, showing the overall results with big marks. Two-phased approaches (ACR+ACEw and ACR+ACEb) have a capability of high recall rate and small data spread, because of the letter-unit based ACR module. And ACEw and ACEb have a capability of high precision rate, which is derived from the use of block-unit/word-unit frequency data. In addition, the block-unit approaches (ACEb and ACR+ACEb) show better results than the word-unit approaches (ACEw and ACR+ACEw). We considered the block-unit frequency data resolved the data sparseness problem of the word-unit approach. Our method is effective for any areas and achieved high precision rate and high recall rate by the combination of these capabilities.

The ACEw result shows the low recall rate. This should be caused by long word of the toponyms of Finland and Germany. They have many long words in the toponyms and they may cause unknown words. On the other hand, ACR+ACEw shows high recall rate, because ACR can reduce the number of area candidates; 1.6 candidates on average in Finnish toponyms and 1.5 candidates on average in German toponyms. This means that ACR plays the main role to perform the toponym resolution and ACE is a module for the precision rate improvement.

5.6 Experiments on Related Areas

Our experiment assumed 1-to-1 relationship between the areas and their languages. But there are many areas which use the same language, for example, Spain and Chile as Spanish-speaking areas or China and Taiwan as Chinese-speaking areas. Areas may contain same or similar toponyms for linguistic or historical reasons, and we need to consider these related areas for a better performance. It can be a problem for the toponym resolution among these areas because of the similarity of surface statistics.

Table 9 Empirical result including related areas.

area	F-measure	Precision (%)	Recall (%)
China	.78 (84.8%)	74.53 (84.1%)	82.50 (85.7%)
Taiwan	.83 (-)	82.29 (-)	83.20 (-)
Japan	.90 (100.0%)	86.99 (100.3%)	93.57 (99.5%)
Thailand	.97 (99.0%)	97.36 (99.0%)	96.56 (99.8%)
Greece	.85 (98.8%)	82.80 (100.0%)	88.30 (99.6%)
Finland	.84 (100.0%)	79.27 (100.0%)	89.34 (99.6%)
Germany	.90 (100.0%)	83.82 (99.9%)	96.48 (99.8%)
France	.80 (98.8%)	76.29 (98.9%)	83.36 (98.8%)
Spain	.75 (88.2%)	71.73 (86.6%)	78.12 (88.9%)
Chile	.80 (-)	81.39 (-)	78.81 (-)
U.S.A.	.92 (100.0%)	91.62 (99.5%)	92.43 (99.6%)
TOTAL	.85 (95.5%)	82.36 (96.3%)	87.52 (95.2%)

Table 9 shows a result of our system with eleven areas, Chile and Taiwan added as related areas of Spain and China to the initial experiment (Sect. 4.1). The numbers shown in parentheses in Table 9 represent the ratio comparing to the result of the initial experiment (Table 8). Table 9 shows that the addition of Chile and Taiwan, only one related area each, decreased the accuracy of Spain and China to about 85% of the initial experiment. Thus we need to consider a language-robust approach to solve this problem. On this issue, Sano et al. proposed another approach targeting these related areas [10].

Our approach is mostly based on *n*-gram frequency, thus it is obvious that our approach is not robust enough against an area set including related areas. In this paper we concentrate on an effective use of prefix-type and suffix-type blocks as a significant surface information of toponyms for a robust toponym resolution. Table 9 also shows that few degradation for all areas except Spain and China. We consider that this shows the robustness of our approach.

6. Conclusion

For a robust toponym resolution, we proposed an automatic two-phased method with block-unit frequency data based only on surface statistics of toponyms, which requires only a simple gazetteer to construct toponym corpora for the training of area-specific statistical information. Our method is an efficient combination of an area candidate reduction module and an area candidate examination module which uses block-unit frequency data to improve the precision. We showed that our method succeeded in obtaining high precision rate, 85.54% on average, along with high recall rate, 91.92% on average. Our method is based on simple statistical information and it is flexible enough to be adapted to any area. And also, it is robust enough to be able to estimate

the area candidates for unknown toponyms.

References

[1] G. Zhou and J. Su, "Named entity recognition using an HMM-based chunk tagger," ACL '02: Proc. 40th Annual Meeting on Association for Computational Linguistics, pp.473–480, 2001.
[2] J.R. Curran and S. Clark, "Language independent NER using a maximum entropy tagger," Proc. CoNLL-2003, pp.164–167, 2003.
[3] A.M. Olligschlaeger and A.G. Hauptmann, "Multimodal information systems and GIS: The informedia digital video library," Proc. 1999 ESRI User Conference, pp.102–106, 1999.
[4] H. Li, R.K. Srihari, C. Niu, and W. Li, "InfoXtract location normalization: A hybrid approach to geographic references in information extraction," Proc. HLT-NAACL 2003 Workshop on Analysis of Geographic References, pp.39–44, 2003.
[5] B. Pouliquen, M. Kimler, R. Steinberger, C. Ignat, T. Oellinger, K. Blackler, F. Fuart, W. Zaghouani, A. Widiger, A.C. Forslund, and C. Best, "Geocoding multilingual texts: Recognition, disambiguation and visualisation," CoRR, vol.abs/cs/0609065, pp.53–58, 2006.
[6] P. Clough, "Extracting metadata for spatially-aware information retrieval on the Internet," Proc. 2005 ACM Workshop on Geographic Information Retrieval (GIR '05), pp.25–30, 2005.
[7] E. Garbin and I. Mani, "Disambiguating toponyms in news," HLT '05: Proc. Conference on Human Language Technology and Empirical Methods in Natural Language Processing, pp.363–370, 2005.
[8] E. Rauch, M. Bukatin, and K. Baker, "A confidence-based framework for disambiguating geographic terms," Proc. HLT-NAACL 2003 Workshop on Analysis of Geographic References, pp.50–54, 2003.
[9] T. Sano, S.H. Nobesawa, and H. Saito, "Automatic country identification of area names based on surface features," Proc. 7th International Symposium on Natural Language Processing (SNLP 2007), pp.13–18, 2007.
[10] T. Sano, S.H. Nobesawa, H. Okamoto, H. Susuki, M. Matsubara, and H. Saito, "Toponym resolution based on surface difference among probable countries," Journal of Natural Language Processing. Under review.

Tomohisa Sano received his M.S. degree in computer science from Keio University, Japan. He is a Ph.D. candidate majoring in computer science at Graduate School of Science and Technology in Keio University.

Shiho Hoshi Nobesawa received her Ph.D. degree in computer science from Keio University. After working with Tokyo University of Science, she joined Faculty of Knowledge Engineering in Musashi Institute of Technology (renamed to Tokyo City University) as a lecturer.

Hiroyuki Okamoto received his B.S. degree and his M.S. degree in computer science from Keio University, Japan. He is a Ph.D. candidate majoring in computer science at Graduate School of Science and Technology in Keio University.

Hiroya Susuki received his B.S. degree and his M.S. degree in computer science from Keio University, Japan. He is now a Ph.D. candidate majoring in computer science at Graduate School of Science and Technology in Keio University.

Masaki Matsubara received his B.S. degree and his M.S. degree in computer science from Keio University, Japan. He is now a Ph.D. candidate majoring in computer science at Graduate School of Science and Technology in Keio University.

Hiroaki Saito is an Associate Professor at Keio University, Japan. Research interest lies in natural language processing and music comprehension. Ph.D.