Acquiring Synonyms from Monolingual Comparable Texts

Mitsuo Shimohata[1] and Eiichiro Sumita[2]

[1] Oki Electric Industry Co., Ltd.,
2-5-7, Honmachi, Chuo-ku, Osaka City, Japan
shimohata363@oki.com
[2] ATR Spoken Language Translation Research Laboratories,
2-2-2 Hikaridai, Keihanna Science City, Kyoto, Japan
eiichiro.sumita@atr.jp

Abstract. This paper presents a method for acquiring synonyms from monolingual comparable text (MCT). MCT denotes a set of monolingual texts whose contents are similar and can be obtained automatically. Our acquisition method takes advantage of a characteristic of MCT that included words and their relations are confined. Our method uses contextual information of surrounding one word on each side of the target words. To improve acquisition precision, *prevention of outside appearance* is used. This method has advantages in that it requires only part-of-speech information and it can acquire infrequent synonyms. We evaluated our method with two kinds of news article data: sentence-aligned parallel texts and document-aligned comparable texts. When applying the former data, our method acquires synonym pairs with 70.0% precision. Re-evaluation of incorrect word pairs with source texts indicates that the method captures the appropriate parts of source texts with 89.5% precision. When applying the latter data, acquisition precision reaches 76.0% in English and 76.3% in Japanese.

1 Introduction

There is a great number of synonyms, which denote a set of words sharing the same meaning, in any natural language. This variety among synonyms causes difficulty in natural language processing applications, such as information retrieval and automatic summarization, because it reduces the coverage of lexical knowledge. Although many manually constructed synonym resources, such as WordNet [4] and Roget's Thesaurus [12], are available, it is widely recognized that these knowledge resources provide only a small coverage of technical terms and cannot keep up with newly coined words.

We propose a method to acquire synonyms from monolingual comparable text (MCT). MCT denotes sets of different texts[1] that share similar contents. MCT are appropriate for synonym acquisition because they share not only many

[1] In this paper, "text" can denote various text chunks, such as documents, articles, and sentences.

R. Dale et al. (Eds.): IJCNLP 2005, LNAI 3651, pp. 233–244, 2005.

synonymous words but also the relations between the words in a each text. Automatic MCT construction can be performed in practice through state-of-the-art clustering techniques [2]. News articles are especially favorable for text clustering since they have both titles and date of publication.

Synonym acquisition is based on a distributional hypothesis that words with similar meanings tend to appear in similar contexts [5]. In this work, we adopt loose contextual information that considers only the surrounding one word from each side of the target words. This narrow condition enables extraction from source texts[2] that have different structures. In addition, we use another constraint, *prevention of outside appearance*, which reduces improper extraction by looking over outside places of other texts. This constraint eliminates many non-synonyms having the same surrounding words by chance. Since our method does not cut off acquired synonyms by frequency, synonyms that appear only once can be captured.

In this paper, we describe related work in Sect. 2. Then, we present our acquisition method in Sect. 3 and describe its evaluation in Sect. 4. In the experiment, we provide a detailed analysis of our method using monolingual parallel texts. Following that, we explain an experiment on automatically constructed MCT data of news articles, and conclude in Sect. 5

2 Related Work

Word Clustering from Non-comparable Text

There have been many studies on computing similarities between words based on their distributional similarity [6,11,7]. The basic idea of the technique is that words sharing a similar characteristic with other entities form a single cluster [9,7]. A characteristic can be determined from relations with other entities, such as document frequency, co-occurrence with other words, and adjectives depending on target nouns.

However, this approach has shortcomings in obtaining synonyms. First, words clustered by this approach involve not only synonyms but also many near-synonyms, hypernyms, and antonyms. It is difficult to distinguish synonyms from other related words [8]. Second, words to be clustered need to have high frequencies to determine similarity, therefore, words appearing only a few times are outside the scope of this approach. These shortcomings are greatly reduced with synonym acquisition from MCT owing to its characteristics.

Lexical Paraphrase Extraction from MCT

Here, we draw comparisons with works sharing the same conditions for acquiring synonyms (lexical paraphrases) from MCT. Barzilay et al. [1] shared the same conditions in that their extraction relies on local context. The difference is that

[2] We call texts that yield synonyms as "source texts."

their method introduces a refinement of contextual conditions for additional improvement, while our method introduces two non-contextual conditions.

Pang et al. [10] built word lattices from MCT, where different word paths that share the same start nodes and end nodes represent paraphrases. Lattices are formed by top-down merging based on structural information. Their method has a remarkable advantage in that synonyms do not need to be surrounded with the same words. On the other hand, their method is not applicable to structurally different MCTs.

Shimohata et al. [13] extracted lexical paraphrases based on the substitution operation of edit operations. Text pairs having more than three edit distances are excluded from extraction. Therefore, their method considers sentential word ordering. Our findings, however, suggest that local contextual information is reliable enough for extracting synonyms.

3 Synonym Acquisition

Synonym extraction relies on word pairs that satisfy the following three constraints: (1) agreement of context words; (2) prevention of outside appearance; and (3) POS agreement. Details of these constraints are described in the following sections. Then, we describe refinement of the extracted noun synonyms in Sect. 3.4.

3.1 Agreement of Context Words

Synonyms in MCTs are considered to have the same context since they generally share the same role. Therefore, agreement of surrounding context is a key feature for synonym extraction. We define contextual information as surrounding one word on each side of the target words. This minimum contextual constraint permits extraction from MCT having different sentence structures.

Figure 1 shows two texts that have different structures. From this text pair, we can obtain the following two word pairs WP-1 and WP-2 with context words (synonym parts are written in bold). These two word pairs placed in different parts would be missed if we used a broader range for contextual information.

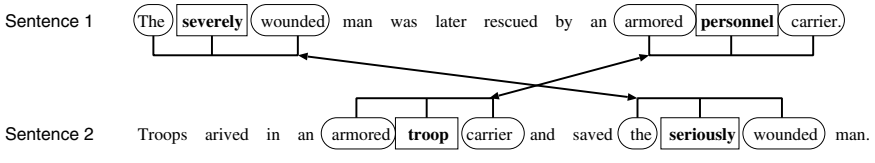

Fig. 1. Extracting Synonyms with Context Words

WP-1 "the **severely** wounded" ⇔ "the **seriously** wounded"
WP-2 "armored **personnel** carrier" ⇔ "armored **troop** carrier"

Words are dealt with based on their appearance, namely, by preserving their capitalization and inflection. Special symbols representing "Start-of-Sentence" and "End-of-Sentence" are attached to sentences. Any contextual words are accepted, but cases in which the surrounding words are both punctuation marks and parentheses/brackets are disregarded.

3.2 Prevention of Outside Appearance

Prevention of outside appearance is a constraint based on characteristics of MCT. It filters incorrect word pairs by looking into outside of synonym words and context words in the other text (we call this outside region the "outside part."). This constraint is based on the assumption that an identical context word — either a noun, verb, adjective, or adverb — appears only once in a text. Actually, our investigation of English texts in the Multiple-Translation Chinese Corpus data (MTCC data described in Sect. 4.1) proves that 95.2% of either nouns, verbs, adjectives, or adverbs follow this assumption.

This constraint eliminates word pairs that have a word satisfying the following two constraints.

C1 The word appears in the outside part of the other text.
C2 The word does not appear in the synonym part of the other text.

The constraint C1 means that the word in the outside part of the other text is considered as a correspondent word, and a captured word is unlikely to be corresponding. In other words, appearance of the word itself is more reliable than local context coincidence. The constraint C2 means that if the word is included in the synonym part of the other text, this word pair is considered to capture a corresponding word independent of the outside part.

Figure 2 illustrates an example of outside appearance. From S1 and S2, the word pair "Monetary Union" and "Finance Minister Engoran" can be extracted. However, the word "Monetary" in S1 does appear in the synonym part of S2 but does appear in another part of S2. This word pair is eliminated due to outside appearance. However, if the word appears in the synonym part of S2, it remains independent of the outside part.

This constraint is a strong filtering tool for reducing incorrect extraction, although it inevitably involves elimination of appropriate word pairs. When applying this constraint to the MTCC data (described in Sect. 4.1), this filtering reduces acquired noun pairs from 9,668 to 2,942 (reduced to 30.4% of non-filtered pairs).

3.3 POS Agreement

Word pairs to be extracted should have the same POS. This is a natural constraint since synonyms described in ordinary dictionaries share the same POS. In addition, we focus our target synonym on content words such as nouns, verbs, adjectives, and adverbs. A definition of each POS is given below.

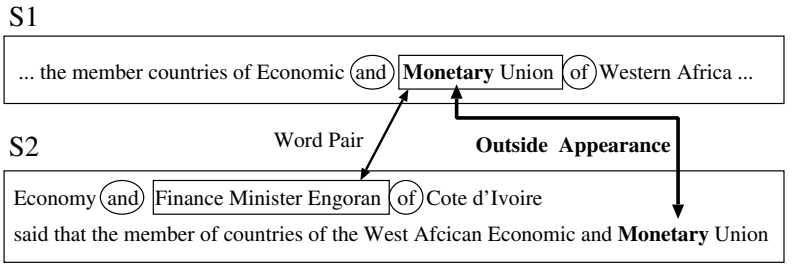

Fig. 2. Text Pair Having Outside Appearance

Nouns	Consist of a noun sequence. Length of sequences is not limited.
Verbs	Consist of one verb.
Adjectives	Consist of one adjective.
Adverbs	Consist of one adverb.

The word pair WP-1 satisfies the constraint for adverbs, and WP-2 satisfies that for nouns. The MCT in Fig. 1 can produce the word pair "the **severely wounded** man" and "the **seriously wounded** man." This word pair is eliminated because the synonym part consists of an adverb and an adjective and does not satisfy the constraint.

3.4 Refinement of Noun Synonym Pairs

Acquired noun pairs require two refinement processes, incorporating context words and eliminating synonyms that are subsets of others, since nouns are allowed to contain more than one word.

After the extraction process, we can obtain noun pairs with their surrounding context words. If these context words are considered to be a part of compound nouns, they are incorporated into the synonym part. A context word attached to the front of the synonym part is incorporated if it is either a noun or an adjective. One attached to the back of the synonym part is incorporated if it is a noun. Thus, when the noun pair "air **strike** operation" = "air **attack** operation" is extracted, both context words remain since they are nouns.

Next, a noun pair included in another noun pair is deleted since the shorter noun pair is considered a part of the longer noun pair. If the following noun pairs Noun-1 and Noun-2 are extracted[3], Noun-1 is deleted by this process.

Noun-1 "British High" ⇔ "British Supreme"
Noun-2 "British High Court" ⇔ "British Supreme Court"

[3] All words in these expressions belong to "proper noun, singular" (represented as NNP in the Penn Treebank manner).

4 Experiment

We used two types of MCT data: sentence-aligned parallel texts (MTCC) and document-aligned comparable texts (Google News). Both data are based on news articles, and their volumes are relatively small. The former data are used for detailed analysis and the latter data are employed to show practical performance. The Google News data consists of both English and Japanese versions. Table 1 shows the statistics of the experimental data, with the major difference between MTCC and Google News data being "Words per Text." The text length of Google News data is much longer than MTCC data since texts in Google News data denote a whole article whereas those in MTCC data denote a sentence.

These two English data and the one Japanese data originally contained plain text data. We applied the Charniak parser [3] to the English data and Chasen[4] to the Japanese data to obtain POS information. It should be noted that we do not use any information except that of POS from parsed results.

Table 1. Statistics of Three Experimental Data

	MTCC	Google News (E)	Google News (J)
Text Clusters	993	61	88
Texts	10,655	394	417
Words	302,474	176,482	127,482
Texts per Cluster (Mean)	10.7	6.5	4.7
Words per Text (Mean)	28.4	447.9	305.7
(Variance)	364.5	64591.3	55495.7

MTCC: Multiple-reference Data from LDC

4.1 Multiple-Translation Chinese Corpus

The Linguistic Data Consortium (LDC) releases several multiple-translation corpora to support the development of automatic means for evaluating translation quality. The Multiple-Translation Chinese Corpus[5] (MTCC) is one of those, and it contains 105 news stories and 993 sentences selected from three sources of journalistic Mandarin Chinese text. Each Chinese sentence was independently translated into 11 English sentences by translation teams. We applied the Charniak parser to these 10,923 translations and obtained 10,655 parsed results. This data comprises high-quality comparable texts, namely parallel texts.

We applied our method to the data and obtained 2,952 noun pairs, 887 verb pairs, 311 adjective pairs, and 92 adverb pairs. Samples of acquired synonyms are shown in Appendix A. Roughly speaking, the number of acquired word pairs for each POS is proportional to the frequency of occurrence for that POS in the MTCC data.

[4] http://chasen.naist.jp/hiki/ChaSen/
[5] Linguistic Data Consortium (LDC) Catalog Number LDC2002T01.

Extracted word pairs were manually evaluated by two methods: evaluation with source texts and without source texts. First, an evaluator judged whether extracted word pairs were synonyms or not without source texts. If two words could be considered synonyms in many cases, they were marked "yes," otherwise "no." The criterion for judgment conformed to that of ordinary dictionaries, i.e., the evaluator judges whether given a word pair would be described as a synonym by an ordinary dictionary. Therefore, word pairs heavily influenced by the source texts are judged as "no," since these word pairs are not synonymous in general situations. Morphological difference (e.g. singular/plural in nouns) is not taken into consideration.

Next, word pairs evaluated as non-synonyms were re-evaluated with their source texts. This evaluation is commonly used in paraphrase evaluation [1,10]. When word pairs could be considered to have the same meaning for the given sentence pair, the evaluator marked "yes," otherwise "no." This evaluation clarifies the ratio of the these two causes of incorrect acquisition.

1. The method captures proper places in sentences from source texts, but the semantic difference between words in this place pair exceeds the range of synonyms.
2. The method captures improper places in sentences from source texts that have the same local context by chance.

An example of evaluation with source texts and without source texts is shown in Fig. 3. Samples of this evaluation are also shown in Appendix A.

The precision, the ratio of "yes" to the total, on MTCC data by each POS is shown in Fig. 4, where the All POS precision with source texts reaches 89.5%. This result suggests that our method could capture proper places of MCT pairs with this level of precision. However, this precision falls to 70.0% without source texts that represents synonym acquisition precision. This is because some of the extracted word pairs have a hypernymous relationship or have great influence on context in source texts.

Acquired word pairs include those occurring only once since our method does not cut off according to word frequency. The amount of those occurring only once accounts for 88.8% of the total. This feature is advantageous for acquiring proper nouns; acquired word pairs including proper nouns account for 63.9% of the total noun pairs.

Word pair judged as non-synonym	
Synonym-1	Muslim robe
Synonym-2	sarong
Source Text Pair	
Sentence-1	A resident named Daxiyate wears a turban and **Muslim robe**.
Sentence-2	A citizen named Daciat wore a Moslem hat and **sarong**.

Fig. 3. Example of Evaluation with Source Texts

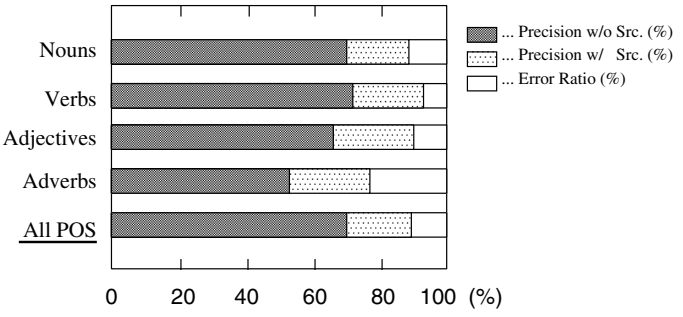

Fig. 4. Precisions for MTCC Data

Here, we discuss our method's coverage of all the synonyms in the training data. Since it is very difficult to list all synonyms appearing in the training data, we substitute identical word pairs for synonym pairs to estimate coverage. We counted identical word pairs from all MCT pairs (Total) and those that have the same context words (Same Context). The ratio of "Same Context" to "Total" denotes coverage of our method and it was found to be 27.7%. If the tendency of local context for identical word pairs is equal to that of synonym word pairs, our method can capture 27.7% of the embedded synonyms in the training data.

We looked up acquired word pairs in WordNet[6], a well-known publicly available thesaurus, to see how much general synonym knowledge is included in the acquired synonyms. We could obtain 1,001 different word pairs of verbs, adjectives, and adverbs after unifying conjugation[7]. WordNet knows, i.e., both words are registered as entries, 951 word pairs (95.0%) among the 1,001 acquired pairs. The thesaurus covers, i.e., both words are registered as synonyms, 205 word pairs (21.6%) among 951 known pairs. This result shows that our method can actually capture general synonym information. The remaining acquired word pairs are still valuable since they include either general knowledge not covered by WordNet or knowledge specific to news articles. For example, extracted synonym pairs, "express"="say," "present"="report," and "decrease"="drop" are found from the data and are not registered as synonyms in WordNet.

4.2 Google News Data

We applied our method to Google News data acquired from "Google News, [8]" provided by Google, Inc. This site provides clustered news articles that describe the same events from among approximately 4,500 news sources worldwide.

[6] http://www.cogsci.princeton.edu/~wn/

[7] Acquired nouns are excluded from the consulting since many proper names are acquired but are not covered in WordNet.

[8] English version: http://news.google.com/
Japanese version: http://news.google.com/nwshp?ned=jp

From the Google News site, we gathered articles with manual layout-level checking. This layout-level checking eliminates unrelated text such as menus and advertisements. Our brief investigation found that clustered articles often have a small overlap in described facts since each news site has its own interest and viewpoint in spite of covering the same topic.

We use entire articles as "texts" and do not employ an automatic sentence segmentation and alignment tool. This is because the results derived from automatic sentence segmentation and alignment on the Google News data would probably be unreliable, since the articles greatly differ in format, style, and content. Since our method considers only one-word-length context in each direction, it can be applied to this rough condition. On the other hand, this condition enables us to acquire synonyms placed at distant places in articles.

The next issue for the experimental conditions is the range for outside-appearance checking. Following the condition of MTCC data, the outside-appearance checking range covers entire texts, i.e., outside appearance should be checked throughout an article. However, this condition is too expensive to follow since text length is much longer than that of MTCC data. We tested various ranges of 0 (no outside-appearance checking), 10, 20, 40, 70, 100, 200, and unlimited words. Figure 5 illustrates the range of outside-appearance checking.

We limit the words to be tested to nouns since the acquired amounts of other POS types are not sufficient. Acquired noun pairs are evaluated without source

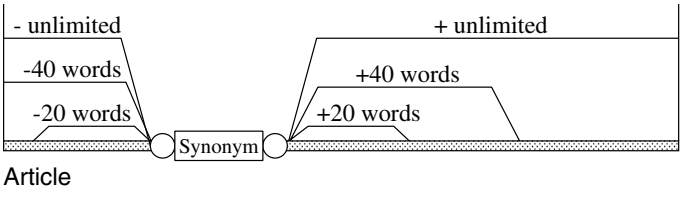

Fig. 5. Range for Outside-Appearance Checking

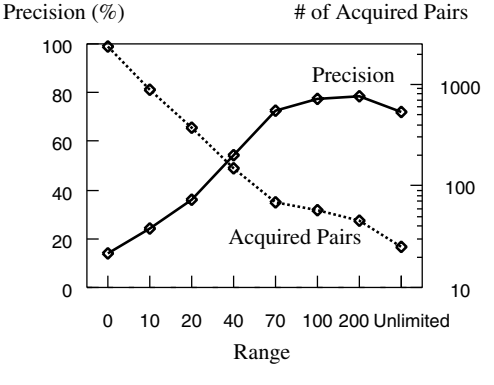

Fig. 6. Precisions of Google (E) by Outside-Appearance Checking Range

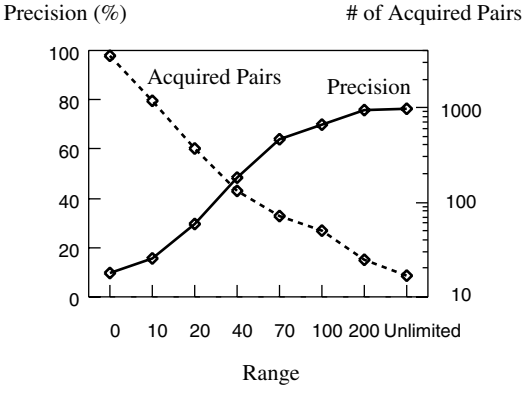

Fig. 7. Precisions of Google (J) by Outside-Appearance Checking Range

texts. Appendix B shows examples. Figures 6 and 7 display the amount and precision for acquired nouns in each range of English data and Japanese data, respectively.

The tendencies of these two data are similar, as the range expands, precision increases and the amount of acquired pairs decreases at an exponential rate. When the range is close to unlimited, precision levels off. The average precision at this stable range is 76.0% in English data and 76.3% in Japanese. The precision improvement (from 13.8% to 76.0% in English data and from 9.5% to 76.3% in Japanese data) shows the great effectiveness of prevention of outside appearance.

5 Conclusions

We proposed a method to acquire synonyms from monolingual comparable texts. MCT data are advantageous for synonym acquisition and can be obtained automatically by a document clustering technique. Our method relies on agreement of local context, i.e., the surrounding one word on each side of the target words, and prevention of outside appearance.

The experiment on monolingual parallel texts demonstrated that the method acquires synonyms with a precision of 70.0%, including infrequent words. Our simple method captures the proper place of MCT text pairs with a precision of 89.5%. The experiment on comparable news data demonstrated the robustness of our method by attaining a precision of 76.0% for English data and 76.3% for Japanese data. In particular, prevention of outside-appearance played an important role by improving the precision greatly.

The combination of our acquisition method, an automatic document clustering technique, and daily updated Web texts enables automatic and continuous synonym acquisition. We believe that the combination will bring great practical benefits to NLP applications.

Acknowledgment

The research reported here was supported in part by a contract with the National Institute of Information and Communications Technology entitled "A study of speech dialogue translation technology based on a large corpus".

References

1. R. Barzilay and K. McKeown. Extracting paraphrases from a parallel corpus. In *Proc. of ACL-01*, pages 50–57, 2001.
2. M.W. Berry, editor. *Survey of Text Mining Clustering, Classification, and Retrieval*. Springer, 2004.
3. E. Charniak. A maximum-entropy-inspired parser. In *Proc. of the 1st Conference of the North American Chapter of the Association for Computational Linguistics*, 2000.
4. C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.
5. Z. Harris. *Mathematical Structures of Language*. Interscience Publishers, 1968.
6. D. Hindle. Noun classification from predicate-argument structures. In *Proc. of ACL-90*, pages 268–275, 1990.
7. D. Lin. Automatic retrieval and clustering of similar words. In *Proc. of COLING-ACL 98*, pages 768–774, 1998.
8. D. Lin, S. Zhao, L. Qin, and M. Zhou. Identifying synonyms among distributionally similar words. In *Proc. of the 18th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1492–1493, 2003.
9. C.D. Manning and H. Schütze, editors. *Foundations of Statistical Natural Language Processing*, pages 265–314. MIT Press, 1999.
10. B. Pang, K. Knight, and D. Marcu. Syntax-based alignment of multiple translations: Extracting paraphrases and generating new sentences. In *Proc. of HLT-NAACL 2003*, pages 181-188, 2003.
11. F. Pereira, N. Tishby, and L. Lee. Distributional clustering of English words. In *Proc. of ACL-93*, pages 183–190, 1993.
12. P.M. Roget. *Roget's International Thesaurus*. Thomas Y. Crowell, 1946.
13. M. Shimohata and E. Sumita. Identifying synonymous expressions from a bilingual corpus for example-based machine translation. In *Proc. of the 19th COLING Workshop on Machine Translation in Asia*, pages 20–25, 2002.

Appendix

A Samples of Acquired Words from MTCC and Their Evaluation

	Synonym-1	Synonym-2	Evaluation
Nouns	press conference	news conference	Yes
	foreign funds	foreign capital	Yes
	complete	finish	Yes
	disclose	reveal	Yes
	military officials	military officers	No
	Sunday radio program	Sunday TV program	No

Verbs	indicate	show	Yes
	believe	think	Yes
	cease	stop	Yes
	consider	study	No
	believe	trust	No
Adjectives	basic	essential	Yes
	notable	significant	Yes
	massive	substantial	Yes
	active	good	No
	direct	strong	No
Adverbs	currently	now	Yes
	certainly	definitely	Yes
	extremely	very	Yes
	now	officially	No
	absolutely	entirely	No

B Samples of Acquired Nouns from Google News (E) and Their Evaluation

	Synonym-1	Synonym-2	Evaluation
Nouns	Karzai	President Karzai	Yes
	Abu Omar	Abu Umar	Yes
	relief effort	relief mission	Yes
	Muslim community	Muslim minority	No
	World Food Program	World Health Organization	No