Estimation of Question Types for Indonesian Question Sentence

Ayu Purwarianti, Tsuchiya Masatoshi, Seiichi Nakagawa
Department of Information and Computer Sciences, Toyohashi University of Technology,
ayu@slp.ics.tut.ac.jp, tsuchiya@imc.tut.ac.jp, nakagawa@slp.ics.tut.ac.jp

Abstract

We propose a question classification system for poor resource language such as Indonesia language. The aim is to provide a good question classification system by only using the available language resource. Our strategy here is to build a shallow parser to extract some important words and make use the results as features for a SVM based question classification. The shallow parser divides question sentence into phrases based on word class of each word. The features for SVM include the shallow parser result, WordNet distance, and word pair's frequency. By using 3000 questions and 6 question types, the highest accuracy score achieved is about 95%.

1. Introduction

Question classification is an important phase in most question answering systems. There are three approaches for question classification[2]: rule based, language modeling and machine learning based. High accuracy score is usually gained by the rule based or machine learning based system. Using a rule based system means that one has to provide many rules to get a high accuracy score. This is not suitable for poor resource language such as Indonesian language. The machine learning approach appears as the solution for the time consuming process problem.

There are many question classification researches with machine learning approach. Many of them utilized some language tools such as named entity tagger, morphological analyzer or sentence parser. On average, the result gained by specific feature is higher than using bag of words for a coarse grained classification.

Zhang[7] compared various machine learning methods for the question classification such as Nearest Neighbor, Naive Bayes, Decision Tree, Sparse Network of Winnows (SnoW) and Support Vector Machine (SVM). By using bag of word and bag of ngram features, the SVM algorithm gave the highest accuracy score. The experiments achieved 90% for coarse classes (6 classes) using tree kernel, and 80.2% for fine classes (50 classes).

Li[2] used the SNoW learning architecture to classify questions into coarse (6 classes) and grained (50 classes) category. The feature types include words, POS tags, chunks, named entities, head chunk (the first noun chunk in the sentence), and semantically related words (words that often occur with a specific question class). The last feature will be extracted for a specific class if there is a word in the question belong to the semantically related word list. For example, if "away" occurs in the sentence then the sensor Rel(class=distance) will be active. The highest score for coarse classes was 91%, and for fine classes was 84.2%.

Skowrow[4] utilized the SVM algorithm with features combined from subordinate word category, question focus and syntactic semantic structure. The subordinate word category is taken from WordNet, for example, "kangaroo" will have "animal" as its subordinate word category. The question focus is extracted from manually listed regular expression. The syntactic semantic structure is used because there are structures for a category that don't show on other category. Using an automatic process, they got 147 structures, providing an additional features for various question. The best result was gained by combined all features (85.6% for fine classes category).

Here, we classified Indonesian questions under coarse classes category using an SVM algorithm. As a poor resource language, there is no available language tools can be used to extract specific information from Indonesian question. For that reason, we built our own shallow parser to extract some main words and used it as the feature of the available SVM software. We also added other information such as WordNet distance and word pair's frequency for features on the SVM engine. The experimental result achieved 95% accuracy score for the combined features.

The rest of the paper is organized as follows: Sec. 2 presents the Indonesian shallow parser includes the characteristics of Indonesian question and the shallow parser strategy; Sec. 3 discusses all features we used for the SVM engine; Sec. 4 describes our experimental data and its result.

2. Indonesian Shallow Parser

2.1 Indonesian Question

We found that there are two main problems in designing an Indonesian question shallow parser. The first one is the OOV (Out of Vocabulary) words. We took word class information from an Indonesian-English dictionary (29,054 words), because we couldn't find other Indonesian lexical resource for the word class list. But even though we already used this dictionary (only to take its word class information), still there are many unlisted words in the dictionary. For example, the word "ibukota" (capital city) in sentence "Apakah ibukota Yunani?" (what is the capital city of Greece?). In order to extract the correct main noun, we should know that the word class for "ibukota" is noun.

Here is another example related with compound word (OOV), "Siapakah juara nomor lari 1.500 meter putra di Kejuaraan Dunia Atletik ke-10?" (Who won the 1500 m run in 10th Athletic World Championship?). Here, "nomor lari" (run) should be treated as a single word, but if we only rely on the Indonesian-English dictionary then "nomor lari" will be separated into two words, "nomor" (number) and "lari" (run), which will be placed in different phrases.

Second problem is the Indonesian question grammar structure. Based on our observation on the 3000 questions collected from Indonesian respondents, we found that there are influences from regional languages (such as Sunda language, Java language, Batak language, etc) to the grammar structure of Indonesian question. For example, "Pada tanggal berapakah, terjadinya penandatanganan MOU antara pemerintah Indonesia dan GAM" (when was the MOU between Indonesia government and GAM being held?). The formal Indonesian question will put the word "terjadinya" (predicate) after the word "penandatanganan MOU" (subject). From the question collection, we observed that a main noun can be located before or after the question word, and it also can be placed near or far from the question word.

2.2 Shallow Parser Strategy

The aim of our shallow parser is to get the question word, one main noun and one main verb from the Indonesian question sentence. We believe that these words are important features in the question classification.

The strategy of the shallow parser is to split the question sentence into phrases and choose the main noun and verb from those phrases. The question sentence splitting is done based on the word class. Therefore the word class is an important point in our shallow parser. As has been mentioned before in Section 2.1, we used the Indonesian-English KEBI dictionary for the word class list. But, there are weaknesses on the dictionary, such as its coverage size and its word class labeling rules.

To overcome these weaknesses, we complemented the dictionary by our manual word list for words other than noun, verb and adjective, such as adverb, preposition, etc. We also used Indonesian corpus to define the word class of an OOV word. By assumption that words other than noun, verb and adjective are already listed, we used the corpus to define whether the word class is a noun, a verb or an adjective. For this, we specified some words that usually appear before noun, verb and adjective. Then, for each OOV word, we calculated the frequency of its pair with the defined words and we attributed the word class based on these frequencies. The phrase is assigned based on the word class. From the phrases, we selected the main noun and main verb by rules.

The example on the shallow parser's result is shown in Figure 1. Based on the word class of each word, the parser assigns a PP for the first phrase, an NP for the second and forth phrases, and a VP for the second phrase. The first phrase is a special phrase because it contains the question word "apa". Usually when a noun is grouped with the question word, then the noun becomes the important noun or question focus. But in this example case, the noun ("nama") is included in a list that we call a stop noun list which can be abandoned. Therefore, the system will search in other NPs to get the important noun. And because the verb in the VP is an active word (it begin with prefix "me"), the system selects the headword of the last NP as the important noun ("ibukota").

Question: Dengan nama apakah, warga Siprus Turki menyebut ibukota Nicosia (With what name, citizen of Siprus Turk call the capital city of Nicosia)
Phrases resulted:
- dengan nama apakah (PP)
- warga Siprus Turki (NP)
- menyebut (VP)
- ibukota Nicosia (NP)

Shallow Parser Result:
- question word: apakah (what)
- main noun: ibukota (capital city)
- main verb: menyebut (call)

Figure 1. Question Example and Its Shallow Parser Result

3. Feature for SVM based Question Classification

The first feature for the question classification is the output of the shallow parser. Basically, it represents important words to define a question class of a question sentence. It includes the question word, the preposition before the question word, the main noun, the main verb and its phrase category.

Other than this feature, we also add two more features developed from the main noun. Those are the WordNet distance and the word pair's frequency between the main noun with some defined previous words, which is discussed in the next section.

3.1 WordNet Distance Feature

WordNet describes the semantic relation among words. Thus, there are many researches using WordNet for the question classification system. But those researches usually used it for English questions. Here, we tried to use WordNet for the Indonesian questions. The complete procedure on using WordNet is as follows:

1. Translate the main noun into English using the Indonesian-English KEBI dictionary.
2. Calculate WordNet depth between the translated noun with some specified WordNet synsets that represent the question category.
3. Include all these WordNet distance as the additional attribute value.

For English question, the strategy using WordNet will improve the accuracy score quite significant. For Indonesian question, this strategy has problems with the translation ambiguation and the OOV words. The translation ambiguation is handled by using all distances in the additional feature, such as mentioned in step 3 above. The OOV words can be categorized into common noun, Indonesian proper name and borrowed words. By not translating the OOV words, WordNet is available for English borrowed words such as "distributor" in sentence "Apa nama distributor rekaman

CD acara festival Raum & Schatten di Berlin, untuk Indonesia?" (What is the name of CD record distributor for Raum & Schatten festival in Berlin?). But this method doesn't work for the common noun and Indonesian proper name.

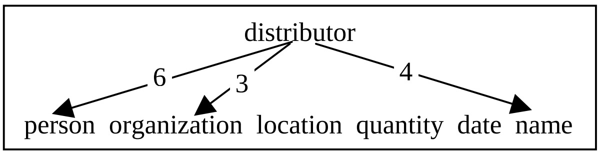

Figure 2. Example on WordNet distance

3.2 Preceding Word Pair's Frequency Feature

The second feature is the word pair's frequency with some defined preceding words. The method is alike with the method to attribute the word class using Indonesian corpus, mentioned in Section 2.2. The different part is the word list and the last process. The overall procedure is as follows:

1. List some words for each question category (person, name, organization, location, date, quantity). For example, words for "person" include president, teacher, musician, doctor, writer, etc.
2. Search in the corpus for some most frequent preceding words of those words in step 1.
3. Select some words from words in step 2 that are significant enough to differentiate question category. This step is done manually and resulted 6 word lists (see Table 1).
4. For each question, calculates the word pair frequency between the main noun in the question and each listed word resulted from step 3.
5. Include all the frequencies as the additional feature. We don't take one highest frequency, instead we include all the frequencies because a noun might have some most frequent preceding words spread among the 6 word lists.

This method is able to handle OOV words such as common noun or proper name that can't be solved by the WordNet distance. For example, the common noun "bandara" in "Apakah nama bandara di Pekanbaru?" (What is the airport located in Pekanbaru?), or the proper name "Biodiesel" in "Apakah nama kimia untuk Biodiesel?" (What is the chemistry name for Biodiesel?). For both nouns ("bandara" and "Biodiesel"), WordNet distance approach couldn't give additional information because these words aren't listed in both the Indonesian-English dictionary and the WordNet itself. But by using the preceding word pair's frequency approach, we still able to gain additional information that can distinguish the semantic information between "bandara" (as location) and "Biodiesel" (as name).

Another advantage is that the data resource needed for this method. For WordNet distance approach, we need a bilingual dictionary and a thesaurus which demand an expensive effort if one is not available. But for the listed preceding word pair's frequency method, it only needs a monolingual corpus that can be collected from the WWW.

Table 1. Word List for Each Question Category

Person	oleh, calon, para, seorang, orang, menjabat, ungkap, ujar, ucap, kata, profesi, kalangan
Organization	anggota, antar, aset, bawah, ketua, kepala, pemimpin, pimpinan, manajemen, milik, kantor, kelompok
Location	kawasan, daerah, dari, arah, dekat, menuju, sekitar, ke, masuk, pembangunan
Quantity	beberapa, puluhan, ratusan, ribuan, jutaan, belasan, milyaran, puluh, ratus, ribu, juta, belas, milyar
Date	akhir, awal, tengah, hingga, ketika, saat, waktu, sejak, selama, setiap, tiap
Name	seekor, sebelum, seluruh, seputar, setelah, sosialisasi, sebatang, sekuntum, terjadi, usai, pelaksanaan, peluncuran, meraih, korban, kotak, sehelai, judul, karangan, hasil

4. Experiments

4.1 Experimental Data

In our knowledge, there is no Indonesian Question Answering data available. For this reason, we built our own Indonesian question-document pairs. We collected Indonesian articles from two popular Indonesian newspaper sites (tempointeraktif.com and kompas.com) for data years 2000 – 2005 (71,109 articles; 23 million sentences). For the question collection, we asked 15 Indonesian people to write Indonesian questions based on 212 articles that we selected manually from the Indonesian corpus. Each respondent wrote factual wh-questions (what, when, where, who and which questions) for 6 question classes (person, organization, location, quantity, date and name). After eliminating the similar question, it gave us 3000 questions, 500 questions for each question category.

We also used some available resources such as WEKA machine learning software (http://www.cs.waikato.ac.nz/ml/weka/), Indonesian-English dictionary (KEBI[14], 29,054 Indonesian words) and WordNet in English.

4.2 Experimental Result

In the experiment, we used an SVM algorithm with linear kernel and the "string to word vector" function to process the string value, both are available in the WEKA software. For the baseline, we used the bag of words feature. As for the machine learning comparison, we tried the baseline feature for some machine learning methods: C4.5, K Nearest Neighbor (kNN) and SVM. The result is in Table 2. The highest score is achieved by using the SVM algorithm.

We compared the highest baseline result with our proposed features (See Table 3. for the result): the shallow parser result (SP), the WordNet distance (WN) and the preceding word pair's frequency (PF). We used 10-fold cross validation for the accuracy calculation.

Table 2. Accuracy Score of Several Machine Learning Algorithms with Bag of Words Feature in the Indonesian Question Classification

Method	Accuracy Score
C4.5	88.5%
K Nearest Neighbor	72.3%
SVM	92.9 %

Table 3. Accuracy Score of Indonesian Question Classification

Method	Accuracy Score
Baseline	92.9%
SP	94.0%
SP + WN	94.7%
SP + PF	95.0%
SP + PF + WN	95.1%

From Table 3., we can see that using only several important words (SP) gives higher score than using all words in the question (baseline). It strengthens the conclusion of other researches on the machine learning based question classification. The result using the word pair's frequency is a bit higher than using WordNet distance. We assume that this is mostly because of the OOV words. And the highest accuracy score is achieved by using all features (SP + PF + WN). It improves the baseline score for about 2.26%.

Table 4. Confusion Matrix for Baseline System

out in	person	org	loc	quan	date	name
person	491	6	0	0	0	3
org	48	397	19	0	0	36
loc	2	17	456	0	0	25
quan	0	0	0	494	6	0
date	0	1	0	2	497	0
name	1	36	9	3	0	451

As shown in Table 4, the highest misclassification for the baseline system is the "organization" category falls into "person" category. This is mostly for "who" question, such as "Didukung oleh siapa sajakah, Soeharto yang terpilih menjadi kepala negara tujuh kali lewat MPR?" (Who supported Soeharto becoming country leader seven times through MPR?). Human without any knowledge of Soeharto and MPR will also have difficulty in deciding the question class. Another high error lies in the misclassification between organization and name class. The result shows that there are 36 organization questions misclassified as name questions, and there are also 36 name questions misclassified as organization questions. These errors mostly happen on "what" and "which" questions.

From Table 5., we can see that using shallow parser gives better result in handling "what" and "which" question. The misclassification on "name" question is significantly decreased from 36 into 13 questions. But unfortunately, the person-org and org-name misclassifications are higher than the baseline result. These errors happen on the "who", "what" and "which" questions.

Table 5. Confusion Matrix for Shallow Parser Feature

out in	person	org	loc	quan	Date	name
person	489	7	0	0	0	4
org	43	386	14	0	0	57
loc	1	9	471	0	0	19
quan	0	0	0	494	6	0
date	0	0	0	1	499	0
name	1	13	4	1	0	481

Based on Table 6., the significant improvement on using additional feature (WN + PF) is on "organization" classification. It improves the shallow parser result on the "organization" class from 77.2% into 83.4% correctness. Compared to the baseline result, the full features gives better accuracy score almost for all classes except the "person" class which is misclassified into "organization" class.

Table 6. Confusion Matrix for SP + WN + PF Feature

out in	person	org	loc	quan	date	name
person	485	11	0	0	0	4
org	42	417	12	0	0	29
loc	1	12	481	0	0	6
quan	0	0	0	495	5	0
date	0	0	0	1	499	0
name	1	18	3	1	0	477

5. Conclusion

Our experiments showed that a shallow parser is able to increase the accuracy score of SVM based question classification using bag of words feature only. By further processing on the shallow parser result, it even got higher accuracy score. Here, we also showed that by using restricted language resource, the question classification using SVM was able to achieve good accuracy score.

Acknowledgement

This work was partially supported by The 21st Century COE Program "Intelligent Human Sensing" and the Hori Information Science Promotion Foundation.

References

[1] KEBI, Kamus Elektronik Bahasa Indonesia, http://nlp.aia.bppt.go.id/kebi/, February 2004.

[2] Kentjono, Djoko, et.al, "Bahasa Indonesia untuk Penutur Asing", Wedatama Widya Sastra, 2004.

[3] Li, Xin and Dan Roth, "Learning Question Classifiers", COLING 2002.

[4] Skowron, Marcin and Kenji Araki, "Effectiveness of Combined Features for Machine Learning Based Question Classification", Journal of Natural Language Processing 2005, pp. 63-83, 2005.

[5] Witten, Ian H. and Eibe Frank, "Data Mining: Practical Machine Learning Tools and Techniques 2nd edition", Elsevier Inc., 2005.

[6] WordNet, http://wordnet.princeton.edu/, February 2004.

[7] Zhang, Dell and Wee Sun Lee, "Question Classification using Support Vector Machine", ACM SIGIR 2003.