What to Classify and How: Experiments in question classification for Japanese

Rebecca Dridan[*]
Computational Linguistics
Saarland University
rdrid@coli.uni-sb.de

Timothy Baldwin
CSSE Department
The University of Melbourne
tim@csse.unimelb.edu.au

Abstract

This paper describes experiments in Japanese question classification, comparing methods based on pattern matching and machine learning. Classification is attempted over named entity taxonomies of various sizes and shapes. Results show that the machine learning based method achieves much better accuracy than pattern matching, even with a relatively small amount of training data. Larger taxonomies lead to lower overall accuracy, but, interestingly, result in higher classification accuracy on key classes.

1 Introduction

Question classification is the primary aim of the question analysis stage of a question answering (QA) system. The goal of a QA system is to return an answer to a question framed in natural language and question classification aims to narrow down the possible types of answer that are likely, given the question. These answer types are commonly called *expected answer types* (EAT) and can range from classifications as broad as PERSON, NUMBER, LOCATION to those as specific as BASEBALL TEAM.

These answer types are generally referred to as named entity (NE) types, although they also include classes for dates and numbers rather than just entities referred to by name. A set of named entity types that is being used for question classification is known as a named entity taxonomy. The taxonomies used in various QA systems vary widely in size, and can be organised as a flat set or in a hierarchy. There are differences of opinion over the optimal size of a NE taxonomy used in question classification. Proponents of small (8–12) type taxonomies (e.g. Kurata et al. 2004) claim that large taxonomies lead to less accurate classification and hence less accurate QA results. Those who use larger sets (e.g. Breck et al. 2001) believe that the inaccuracies in classification are offset by the discriminative power of more types.

Methods for determining these EATs can be classified (as is common in many NLP tasks) as statistical (machine learning) or symbolic (most often pattern matching). Li and Roth (2006) showed that machine learning was a very effective method for question classification, using a test set of 1000 English questions. Despite these results, less than a quarter of the systems at the most recent QA evaluation forums (TREC 2006, CLEF 2006 and NTCIR 5) used machine learning to find EATs. One reason for this might be a perceived lack of training data. The experiment in Li and Roth (2006) used 21,500 English questions annotated with answer type; often this quantity of data is just not available, particularly for languages other than English. One goal then of the experiments reported here is to investigate whether machine learning is viable for question classification when much less data is available, in this case 2000 annotated Japanese questions. The second aim is to see how accuracy is affected by the size of the NE taxonomy.

2 The Data

2.1 Question Sets

The data set used for development is a set of 2000 Japanese questions described in Sekine et al. (2002b). Each question has been annotated with question type (WHO, WHAT, WHEN, ...) as well as the EAT taken from Sekine's Extended Named Entity (ENE) hierarchy (Sekine et al., 2002a). A second smaller set of 100 questions taken from the QAC-1 evaluation (Fukumoto et al., 2002) was manually annotated as a held out test set.

[*]This research was carried out while the first author was a student at The University of Melbourne.

	Size	Flat?	Comment
NE4	4	Yes	based on Li and Roth's coarse grained taxonomy
NE5	5	Yes	NE4 with ORGANIZATION added
IREX	9	Yes	standard IREX supplemented with a NUMERIC class
NE15	15	Yes	drawn from top two levels of ENE
NE28	28	No	augmented NE15, hierarchically splitting NUMBER, TIME and TIME PERIOD
ENE	148	No	version 4 of Sekine's ENE

Table 1: Summary of the characteristics of each Named Entity taxonomy used for question classification in these experiments

2.2 Named Entity Taxonomies

A set of six named entity taxonomies was defined; this was designed to investigate the effects of using taxonomies of different sizes and shapes. As the question data is annotated with types from Sekine's ENE hierarchy, all taxonomies are based on this hierarchy, and were created by merging various types into set categories. Two of the larger taxonomies are hierarchical; this allows us to investigate how the accuracy is affected if soft decision making is used later in the question answering process. Soft decision making in this instance means that if the predicted EAT does not match an answer candidate, then types around that EAT in the hierarchy may be considered, according to a defined strategy. We use a *most likely type* strategy in evaluation, where the most likely type is selected during classification, but direct ancestors or siblings of that type are also accepted. Other strategies might involve selecting the type that is the lowest common parent of likely types. The six taxonomies are described below, with their vital statistics summarised in Table 1.

The experiments in Li and Roth (2006) used both a coarse and a fine-grained taxonomy. In order to make some rough comparisons with this work, we have tried to approximate their coarse-grained set, which consisted of ABBREVIATION, ENTITY, DESCRIPTION, HUMAN, LOCATION and NUMERIC. None of the types in the ENE correspond to ABBREVIATION or DESCRIPTION, so we have defined two small sets, NE4 and NE5, where NE4 consists of THING, PERSON, LOCATION and NUMBER, and NE5 adds an ORGANIZATION class. In NE4, organisations are considered part of the PERSON class. Both of these sets are flat taxonomies.

The IREX named entity set (ORGANIZATION, PERSON, LOCATION, ARTIFACT, DATE, TIME, MONEY and PERCENT) was originally defined for an NE extraction task in the Japanese IREX project in 1998 (Sekine and Isahara, 1999). It has since been one of the standard NE type sets used in systems at the NTCIR workshops. We have defined a flat taxonomy of 9 types, based on the IREX set, but with an added NUMBER class for any numbers not related to money or percentages.

The taxonomies NE15 and NE28 are mid-range taxonomies that allow more expressive type labels, without being as fine-grained as the full ENE. NE15 is selected mostly from the top and second layer of ENE and is a flat taxonomy. NE28 expands on NE15 by allowing further divisions of NUMBER, TIME and TIME PERIOD, in a hierarchical manner. The sixth taxonomy used is the full ENE hierarchy, version 4, which is hierarchical and contains 148 types.

3 Pattern Matching

Rule based classifiers based on pattern matching were the most common question classification technique used at all the recent QA evaluations. To investigate both the issues involved and the accuracy achievable with this technique, the first question classification experiment used a set of manually created pattern matching rules for Japanese.

This task highlighted some of the issues that are peculiar to Japanese NLP. The first issue occurs because Japanese, like Chinese and Thai, is a non-segmenting language and hence questions need to be tokenised into 'words' before further processing. Various tokenisation tools exist for Japanese but there is no definition for the ideal lexical unit and so what constitutes a token varies between tools. Some Japanese processing systems split text into *bunsetsu*, which are logically larger than English words, similar to phrase chunks, while most of the standard tokenisers use smaller units, typically morphemes. Even within the morpheme based tools, tokenisation is inconsistent and it is necessary to use the same tokenisation tool for a whole system to avoid the systematic differences that occur when using different tokenisers. In all of the following experiments, we use ChaSen (Matsumoto et al., 1999) for tokenisation.

Other issues arise from the characters used to write Japanese. Japanese can be written using

4 different scripts—kanji, hiragana, katakana and the Latin alphabet. Kanji are Chinese characters which can have multiple pronunciations and are logographic in nature. There are thousands of kanji characters, although only 2000–3000 are commonly used. Hiragana and katakana are both phonetic scripts with 46 basic characters each. Hiragana is used for inflected verb endings, particles and other words where the kanji is either not known, or too formal. Katakana, on the other hand, is generally only used for foreign loanwords, or sometimes (e.g. in advertising) for emphasis. The Latin alphabet is often used in Japanese, particularly for acronyms. While there are kanji representations for numbers, the Arabic numbers are also often used.

All these scripts create two main problems – the first is that a single word can be written in different alphabets and so word based pattern matching rules have to accommodate this. The second problem is that there are too many characters to be represented by the ASCII character set. Different encoding systems exist that can represent Japanese, including UTF-8, which is slowly becoming an international standard and EUC-JP, which is the most common encoding system in Japan. While it is possible to convert from one encoding to another, this adds an extra layer of complexity and requires knowing in advance what encoding is used in a particular data set or tool. All of these issues affected the design of the pattern matching rules used, and so while the logic of the rules is very simple, the rules are complex and took many days to construct. The rules also ended up being very specific to the encoding system and the tokeniser used.

The basic logic of the rules first determines which one of 16 *question types* (e.g. 誰 *dare* "who", いくら *ikura* "how much", 何時 *nanji* "what time") applies to a question, by matching different variations of question words. Questions that contain 何 *nani* "what" are further processed so that fragments such as "what place" are classified as WHERE. Using this method question type can be determined with an accuracy of 96-99%. An abbreviated example of a rule for determining question type is shown in Algorithm 1.

Once the question type is determined, different rule sets are used depending on this type. Most of these rules looked for a word at a particular distance from the question word (called here the question focus), and checked whether this word was a recognised named entity type. Algorithm 2 shows one such rule. In addition, a small lookup table was hand-constructed that associated common synonyms for named entity types with their type. Some examples from the table are in Figure 1. It was difficult to construct such a set that improved accuracy but was not over-fitted to the data, and so the set remained small, containing only very specific terms.

Algorithm 1 Question type rule example

if *word*[*i*] starts with 何 or *word*[*i*] =なに or *word*[*i*] =なん or *word*[*i*] = なんと or *word*[*i*] = なんで **then**
  if *word*[*i*+1] = 処 or *word*[*i*+1] = 所 **then**
    return doko
  else if *word*[*i*+1] = 年ぶり or *word*[*i*+1] = 年度 or *word*[*i*+1] = 年 or *word*[*i*+1] = 日 or *word*[*i*+1] = 年前 or *word*[*i*+1] contains 月 or *word*[*i*+1] = ねん **then**
    return nannen
  else if *word*[*i*+1] = ドル or *word*[*i*+1] = 円 **then**
    return ikura
  else if ... **then**
    ...
  else
    return nani
  end if
end if

Algorithm 2 Answer type rule example

if *qtype* = *dono* **then**
  qfocus ← *qtype_index*+1
  if *word[qfocus]* in *lookuptable* **then**
    qfocus ← lookup(*qfocus*)
  end if
  if *qfocus* in *taxonomy* **then**
    return *qfocus*
  end if
  return country ▷ default for this type
end if

If the question focus could not be determined, the EAT was classified as the most likely type for that question type, given none of the previous rules applied. Results for this classification method are described in §5.

Canonical	Alternative
テレビ番組名	番組名
ポイント	点数
物質名	ミネラル
音楽名	曲

Figure 1: Lookup table examples: includes abbreviations, alternative orthography and synonyms.

4 Machine Learning

The learner used for the machine learning classification experiments was TiMBL version 5.1, a memory based learner (Daelemans et al., 2004). TiMBL classifiers can be trained very quickly, and hence allow a fast turnaround when evaluating the effects of different features. The learning algorithm used was TiMBL's IB1 algorithm, with $k = 1$ and features weighted by their Gain Ratio value. While some of the features (detailed later) were binary, many were symbolic, such as words or named entity types. In order to use the information that, for example, `country` is more similar to `province` than `person`, the modified value difference metric (MVDM) was used as a similarity measure between feature values.

The classifiers were evaluated first using a leave-one-out cross-validation strategy (train on 1999, test on one), for each of the 2000 questions in the development set. Next, to see how well the training generalised to unseen data, a separate evaluation was run on the 100 question held out set, using the 2000 questions as training data.

4.1 Features

The feature sets all had the same basic format: presence or absence of each of a list of question words, and then context information for each question word (if it appeared). The list of question words was built from the keywords used to identify question types in the pattern matching experiment. This produced 11 words, each of which could have some orthographic variation.

Two content words from the start and from the end of each question were also used as features. Müller (2004) has documented how words at the start of a question are important in question classification, but this work only analysed English questions. In our development set, the average position of question words was the sixth morpheme from the end. 46% of questions had the question word within four morphemes from the end. Importantly, about 5% of the questions occured without a question word, many were fragments, such as:

センデロ	ルミノソ	の
sendero	*ruminoso*	*no*
Sendero	Luminoso	GEN

日本語訳	は	？
nihongoyaku	*ha*	*?*
Japanese translation	TOP	?

"The Japanese translation of Sendero Luminoso is?"

In questions of this form, the words immediately before the topic marker are generally indicative of the EAT.

Both Li and Roth (2006) and Kimura et al. (2005) emphasised the importance of using features that encode linguistic knowledge, especially when the amount of training data is small. The easiest linguistic information to add to the data was the base form of inflected verbs and part of speech tags which were both available from ChaSen. In this experiment, the part of speech was primarily used to filter out non-content words like case markers and verb endings.

The four feature sets used are outlined below, with Feature Set 1 being designed as a baseline experiment, to be directly comparable to the pattern matching experiment since it uses the same information. The other three sets use the POS information from ChaSen to filter out non-content words and then attempt to add semantic information, as used in the Li and Roth (2006) experiment, using whatever linguistic resources were available. Li and Roth used 4 different sources of semantic information: named entity tagging, WordNet, class-specific related words and distributional similarity based categories. In the experiments here Feature Sets 2 and 3 relate to named entity tagging, while Feature Set 4 uses a similar information source to WordNet. The class-specific related words appear to be similar to the lookup table used in the pattern matching experiment, but resources were insufficient to create a similar list large enough to be useful, nor was there a pre-compiled distributional similarity list for Japanese.

Feature Set 1: Word based only

The first set of features used were based on words only (or more precisely, on morphemes according to the ChaSen definition), and hence could be considered to use similar information to that available in the pattern matching experiment. As well as the binary features indicating the presence

ushikubo	*takio*	*san*	*ha*	*genzai*
Ushikubo	Takio	TITLE-HON	TOP	current

saitamaken	*no*	*doko*	*ni*	*dōjō*	*wo*
Saitama	GEN	where	DAT	dojo	ACC

hirai	*te*	*iru*	*ka*
is operating			QM

"Where in Saitama is Takio Ushikubo currently operating his dojo?"

	Set 1	Set 2
first word:	*ushikubo*	*ushikubo*
second word:	*takio*	*takio*
last word:	*iru*	*hiraku*
second last word:	*te*	*dōjō*
doko **present:**	yes	yes
word *doko* **- 1:**	*no*	`location_rel`
word *doko - 2:*	*saitamaken*	*genzai*
word *doko + 1:*	*ni*	*dōjō*
word *doko + 2:*	*dōjō*	*hiraku*

	Set 3	Set 4
first word:	*ushikubo*	*ushikubo*
second word:	*takio*	*takio*
last word:	*hiraku*	*hiraku*
second last word:	*dōjō*	*dōjō*
doko **present:**	yes	yes
word *doko* **- 1:**	`location_rel`	`location_rel`
type *doko* **- 1:**	PROVINCE	-
word *doko* **- 2:**	*genzai*	*genzai*
type *doko* **- 2:**	-	TIME
word *doko* **+ 1:**	*dōjō*	*dōjō*
word *doko* **+ 2:**	*hiraku*	*hiraku*

Figure 2: An example sentence and the features that would be instantiated for it.

or absence of each question word, the two words either side of each question word were added if the question word was present, as well as the two words at each end of the question (ignoring the question marker if it was present). This gave a total of 59 features, although any one question would normally only have up to 9 instantiated (if only one question word was present).

Feature Set 2: Using SProUT tags

In order to add more semantic knowledge to the data, SProUT (Becker et al., 2002) was used to tag named entities within the question. SProUT uses a fairly small tag set, predominantly tagging names as `person_rel`, `organization_rel` or `location_rel`, and specifying location types where known. A new feature set was created, again recording the presence of question words, and their surrounding context, but if any of the surrounding words had been tagged by SProUT, the SProUT tag was used instead of the word. This has the effect of making, for example, all countries look the same to the classifier, increasing feature counts while ignoring irrelevant information (such as which country).

Feature Set 3: Using Sekine's Named Entity list

A second semantic information source was the gazetteer style list provided with the ENE hierarchy. This list had over 60,000 words and their associated NE type from the taxonomy. For every feature related to a context word or tag in Feature Set 2, another feature was added to record the type of that word if it occurred in the list. That meant the possible number of features went from 59 to 107. Looking through the data, these features were instantiated most often for the names of countries and also for position titles (e.g. President). This feature added more specific named entity information than the SProUT tags, in a form directly related to the possible EATs.

Feature Set 4: Using an Ontology

WordNet (Miller et al., 1990) has often been used as a source of semantic information in English. There is no Japanese version of WordNet, but the Hinoki ontology was available. This ontology was semi-automatically constructed from a Japanese dictionary as described by Bond et al. (2004) and has a fairly limited coverage of about 30,000 words, with very few proper names. This was used to look up each context word in the feature set to see if it was related to any of the named entity types in the ENE. If a related named entity type was found for any word, it was added as a feature, in the same way that NE types were added in Feature Set 3.

An example of how these features were instantiated for a particular sentence is shown in Figure 2.

5 Results

The results for each experiment across all the NE taxonomies are shown in Table 2. These results show that for every NE taxonomy, machine learning techniques were more accurate, with the most accurate results significantly higher than pattern matching for all taxonomies, except NE5. Surprisingly, Feature Set 1, which was based only on words, achieved the best results for the smaller

taxonomies, despite using the same information as the pattern matching. It would be interesting to repeat this experiment, but using POS tags to remove non-content words and see if that gets an even larger improvement. Adding semantic information did not give the expected performance increase, though we begin to see small gains in accuracy due to this extra information as the taxonomies get larger.

While the results show the expected decrease in accuracy as the taxonomies get larger, it is interesting to examine the results in more detail. Breaking down the results by taxonomy, we looked at the classes that were most accurately classified, and those that were difficult to classify. For the NE4 taxonomy the easiest class to classify was NUMBER, correctly identified 94% of the time. THING on the other hand was often misclassified as PERSON, and only correctly labelled in 55% of cases. NE5 differed from NE4 by splitting the PERSON category into PERSON and ORGANIZATION which had the effect of increasing classification accuracy on PERSON, probably because it was now a more tightly defined class. The ORGANIZATION class, however, had very low accuracy (50%), most frequently being misclassified as THING, but also often as LOCATION. This reflects the different ways organisations such as companies can be referred to—in some cases they are places of action and in others, agents of action.

The IREX taxonomy is similar to NE5, but with NUMBER split into DATE, TIME, MONEY, PERCENT and NUMBER. While ORGANIZATION and THING are still mis-labelled in almost half the instances, interestingly, the accuracy for the PERSON and LOCATION classes goes up from the NE5 results. It appears that despite these classes not changing in any way, the feature sets are becoming more discriminatory with the larger taxonomy. Looking at the new numeric classes, DATE and MONEY were the most reliably identified (89% and 94% of instances respectively), but surprisingly PERCENT was only correctly classified 60% of the time, most commonly being mis-labelled as NUMBER. Looking back at the original features, it appears that SProUT tags 何% *nanpāsento* "what percent" as `percentage_rel`, hence removing the question word 何 *nan* "what" which would otherwise be used as a feature to direct attention to the %. This is a peculiarity of using SProUT that would need to be addressed in future experiments.

Like IREX, NE15 also has a PERSON, LOCATION and ORGANIZATION class. The differences are that the number classes have been split differently, into NUMBER, TIME and TIME PERIOD, and that THING has been replaced by a number of more specific classes. Classification accuracy for PERSON, LOCATION and ORGANIZATION is almost exactly the same for IREX and NE15. For the number related classes, TIME and NUMBER are detected with accuracies of 86% and 88% respectively. TIME PERIOD is much harder to identify (53% of instances). It is often mis-classified as either TIME or NUMBER. None of the other classes are correctly identified with any great accuracy. Given that THING was only classified with an accuracy of 54%, it is not surprising that subdivisions of THING are not easily identified. Many of the classes have only a few instances in the data set (only EVENT and PRODUCT having over 20) so there are insufficient training examples to allow reliable classification.

NE28 is the hierarchical version of NE15, allowing more specific classes of TIME, NUMBER and TIME PERIOD while still using the more general classes when appropriate. The other differences involve splitting POSITION TITLE out from PERSON, and FACILITY from ORGANIZATION. FACILITY (which is used for things like *school* and *library*) was shown to be very difficult to identify because, like ORGANIZATION, things of type FACILITY can refer to places or entities, and so they were often mis-classified as LOCATION or ORGANIZATION. Most of the number classes were detected with a high accuracy, suggesting that adding the extra specificity for sub-types of NUMBER is a worthwhile modification. The classification accuracy of PERSON was higher again than NE15, possibly because removing the POSITION TITLE type entities tightened the class definition even further.

When the ENE taxonomy was used for classification, 116 out of 148 types were represented, with only 37 types appearing in the gold standard more than 10 times. Not surprisingly, the classification accuracy over the rare types was generally very low, although many of the numeric types were reliably classified despite very little representation in the training data. In general, numbers were used in very specific patterns and this provides good evidence for classification. Examining the types that did occur frequently, it was interesting to note that PERSON was clas-

	NE4	NE5	IREX	NE15	NE28	ENE
P	0.74	0.73	0.71	0.65	0.57	0.41
1	**0.78**	**0.76**	0.73	0.68	0.62	0.48
2	0.77	0.75	0.73	0.68	0.63	0.49
3	0.77	0.75	0.73	0.68	0.63	0.49
4	0.77	0.75	**0.73**	**0.69**	**0.64**	**0.50**

(a) 2000 question set with leave-one-out cross-validation

	NE4	NE5	IREX	NE15	NE28	ENE
P	0.75	0.75	0.69	0.63	0.57	0.42
1	0.73	0.74	0.71	**0.68**	**0.62**	**0.48**
2	0.77	0.75	0.69	0.64	0.61	0.45
3	**0.78**	**0.76**	0.69	0.66	0.60	0.45
4	0.77	**0.76**	**0.72**	0.64	0.61	0.41

(b) Test set from QAC-1

Table 2: Classification accuracy. The first row from both tables has the results from the pattern matching experiment, and the last four show the results from the four feature sets used in machine learning.

	NE28		ENE	
	Exact	Lenient	Exact	Lenient
P	0.57	0.86	0.41	0.62
1	0.62	0.88	0.48	0.60
2	0.63	0.86	0.49	0.60
3	0.63	0.86	0.49	0.59
4	0.64	0.88	0.50	0.60

(a) Leave-one-out cross-validation

	NE28		ENE	
	Exact	Lenient	Exact	Lenient
P	0.57	0.82	0.42	0.67
1	0.62	0.82	0.48	0.70
2	0.61	0.83	0.45	0.64
3	0.60	0.86	0.45	0.65
4	0.61	0.82	0.41	0.60

(b) Test set from QAC-1

Table 3: Comparing exact and lenient evaluations for question classification over the hierarchical taxonomies.

sified more accurately with this taxonomy than any other. The class was more tightly defined in ENE than in some of the smaller taxonomies, but even compared to NE28, which labelled exactly the same entities PERSON, the classification accuracy was 1.5% higher (92.5%). The other classes that were reliably classified were MONEY, many of the time classes, and COUNTRY. The classes that appeared frequently but were often misclassified were generally sub-classes of ORGANIZATION such as COMPANY and GROUP.

5.1 Lenient Evaluation

To look at the effects of the hierarchical taxonomies, a lenient evaluation was performed on the classifications over NE28 and ENE, using the *most likely type* soft decision making strategy described in §2.2. This demonstrates the effective EAT classification accuracy when soft decision making is used in answer extraction. The results, presented in Table 3, show that under lenient evaluation, while ENE accuracy is still quite low, NE28 is very good, better than the accuracy achieved for even the smallest taxonomy.

6 Discussion

A direct comparison with other results is not possible, since published question classification results from this data set could not be found. The Li and Roth (2006) work in question classification by machine learning is the most similar experiment reported, using similar methods but differing in using a different (English) data set, and different taxonomies. Comparing their coarse grained classification results with the accuracy of NE4 and NE5 classifications achieved here, they demonstrated significantly higher results with 92.5%, compared to 78.2% and 75.5% respectively. Taking the NE28 taxonomy as the closest to their fine-grained taxonomy, the performance difference was larger with NE28 classification accuracy at 64.0% being significantly lower than their best result of 89.3%. While the level of difficulty of the two tasks may differ (since the Japanese questions were much longer on average than the English questions), this would not explain such a large performance difference.

There appear to be two main factors that differentiated the experiments. The first obvious difference is the amount of training data. The Li and Roth results reported here were achieved using

21,500 questions for training. In their experiments exploring the effects of training data set size, they found that by reducing the size of the training set to 2000 questions (the amount used here), they lost between 14 and 18% accuracy. While this partially explains the lower results in these experiments, one of the ideas being explored here is the validity of machine learning with small amounts of training data, and hence we are also interested in any other factor that may have led to higher accuracy.

Looking at Li and Roth's detailed evaluation, their most beneficial types of semantic information were the class-specific related word list and the distributional similarity lists, the two forms of information not used here. Their best combination of semantic features (which was the set without WordNet) achieved a 5% accuracy increase over the combination of WordNet and named entity features. It is not clear how the class-specific related word lists were created, except that they were constructed after the manual analysis of 5000 questions. Experiences during the pattern matching experiment suggest that this sort of resource is difficult to build in a general way without overfitting to the questions being analysed. It would be interesting to see whether a translation of Li and Roth's lists would be beneficial to classification on this Japanese question set, or whether they are too closely tied to the set they were created from. Distributional similarity lists however could be built automatically from, for example, the Mainichi newspaper corpus, and could be expected, from Li and Roth's results, to deliver more accurate results.

7 Conclusion

Pattern matching techniques appear to be the basis for the majority of question analysis modules at recent evaluation forums. This is surprising, given that the results here showed that a very basic machine learning based method produced much better classification accuracy with much less effort. Moreover, while pattern matching methods are very language specific (and in our case specific even to the character encoding and tokenisation tool), the machine learning method could be easily used with a new language, provided data was available.

Looking at the results across the different sized taxonomies, predictably the accuracy decreases as the taxonomy size increases. However, looking at the accuracy for specific classes, the larger taxonomies actually led to greater accuracy on some of the more common classes, such as PERSON and COUNTRY, and also on most of the numeric classes, even the less common ones. This may lead to more useful results than high accuracy over very generic classes. Indeed, in further experiments with these taxonomies (reported in Dridan (2007)), the taxonomy that gave the best results for the QA task was the full ENE, even with the classification accuracies reported here.

This raises the question of why we are classifying these EATs, and what are useful distinctions to make for a question answering system. The greatest contribution to inaccuracies in the larger taxonomies was the vast number of infrequent categories that in smaller taxonomies were all classed together as THING. These classes in the ENE appear to be somewhat arbitrary, a criticism that can also be directed towards the fine-grained taxonomy used in Li and Roth (2006). There are no apparent reasons for NATIONALITY being a subclass of ORGANIZATION but RELIGION a top level class of its own, for example, or, in the Li and Roth taxonomy, why MOUNTAIN is a class, but RIVER is not. Often, the classes in fine-grained NE taxonomies appear to be selected for the purpose of classifying questions in the question set they are being used with.

While it might seem a strange argument to suggest, for example, that we use more fine grained numeric classes just because these are the ones we can easily classify, it makes sense to use whatever information is available. Numbers are generally used in very specific patterns in text, and often asked about in very discriminative ways, and these facts should be utilised in an information seeking application. Equally, if it is difficult to classify questions asking about THINGs, there may not be any point to looking for more specific subclasses of THING. This decision will depend on the task at hand, and there may be certain classes of artifact that are important to the application, but arbitrarily creating many subclasses of THING because they were manually identified in a development set is not necessarily productive.

The conclusion to take from this is that large taxonomies can be used for question classification, but more work needs to be put in to determine which type distinctions will actually be use-

ful for question answering, and how they can be arranged in a hierarchy. If different types of numeric data can be accurately classified and detected, for example, then expanding the number of numeric types in a taxonomy is beneficial. This is an interesting area for future research.

References

Markus Becker, Witold Drozdzynski, Hans-Ulrich Krieger, Jakub Piskorski, Ulrich Schäfer, and Feiyu Xu. 2002. SProUT - Shallow Processing with Typed Feature Structures and Unification. In *Proceedings of the International Conference on NLP (ICON 2002)*. Mumbai, India.

Francis Bond, Eric Nichols, Sanae Fujita, and Takaaki Tanaka. 2004. Acquiring an ontology for a fundamental vocabulary. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages pp 1319–1325. Geneva, Switzerland.

Eric Breck, Marc Light, Gideon S. Mann, Ellen Riloff, Brianne Brown, Pranav Anand, Mats Rooth, and Michael Thelen. 2001. Looking under the hood: Tools for diagnosing your question answering engine. In *Proceedings of theACL-2001 Workshop on Open-Domain Question Answering*, pages pp 1–8. Toulouse, France.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2004. TiMBL: Tilburg Memory-Based Learner, version 5.1, Reference Guide. Technical Report ILK 04-02, Tilburg University, Tilburg, The Netherlands. URL `http://ilk.uvt.nl/downloads/pub/papers/ilk0402.pdf`.

Rebecca Dridan. 2007. *Using Minimal Recursion Semantics in Japanese Question Answering*. Master's thesis, The University of Melbourne.

Jun'ichi Fukumoto, Tsuneaki Kato, and Fumito Masui. 2002. Question Answering Challenge (QAC-1): An Evaluation of Question Answering Task at NTCIR Workshop 3. In *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, pages pp 77–86. Tokyo, Japan.

Yasumoto Kimura, Kinji Ishida, Hirotaka Imaoka, Fumito Masui, Marcin Skowron, Rafal Rzepka, and Kenji Araki. 2005. Three systems and one verifier - HOKUM's participation in QAC3 of NTCIR-5. In *Proceedings of the Fifth NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering and Cross-Lingual Information Access*. Tokyo, Japan.

Gakuto Kurata, Naoaki Okazki, and Mitsuru Ishizuka. 2004. GDQA: Graph driven question answering system - NTCIR-4 QAC2 Experiments. In *Working Notes of the Fourth NTCIR Workshop Meeting*, pages pp 338–344. Tokyo, Japan.

Xin Li and Dan Roth. 2006. Learning question classifiers: the role of semantic information. *Natural Language Engineering*, 12(3):pp 209–228. URL `doi:10.1017/S1351324905003955`.

Yuuji Matsumoto, Akira Kitauchi, Tatsu Yamashita, and Yoshitake Hirano. 1999. Japanese morphological analysis system ChaSen version 2.0 manual. Technical Report NAIST-IS-TR99009, NAIST, Nara, Japan.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, vol 3(no 4):pp 235–244. URL `http://www.cogsci.princeton.edu/~wn/`.

Karin Müller. 2004. Semi-automatic construction of a question treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004)*. Lisbon, Portugal.

Satoshi Sekine and Hitoshi Isahara. 1999. IREX project overview. In *Proceedings of the IREX Workshop*, pages pp. 7–12. Tokyo, Japan.

Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002a. Extended named entity hierarchy. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, pages pp 1818–1824. Las Palmas, Spain.

Satoshi Sekine, Kiyoshi Sudo, Yusuke Shinyama, Chikashi Nobata, Kiyotaka Uchimoto, and Hitoshi Isahara. 2002b. NYU/CRL QA system, QAC question analysis and CRL QA data. In *Proceedings of the Third NTCIR Workshop on Research in Information Retrieval, Automatic Text Summarization and Question Answering*, pages pp 79–86. Tokyo, Japan.