Question Classification for the Icelandic Language: First Steps

Ólafur Páll Geirsson
Reykjavik University / Menntavegur 1, 101 Reykjavik

Abstract

Question classification is important for question answering. This report presents our work on automatic question classification through machine learning approaches for the Icelandic language. We have built a corpus, Gettu Betur, of annotated questions on which we trained a maximum entropy classifier using several features. Our experimental results show that this approach is promising and achieves up to 92% classification accuracy on coarse classes and 85% on fine classes, even beating some state-of-the-art classifiers. We look at how the structure of the Gettu Betur corpus might explain these surprisingly high figures.

1 Introduction

With the gradual increase of critical information being stored in natural language text, the need for an automated question answering system to extract that information becomes more apparent. Such a system would allow the user to ask a question using natural language and receive an accurate answer both quickly and reliably. Current search engines such as Google or Bing may be successful at directing users towards finding an answer from a ranked list of documents. However, they require the user to look through large amounts of text. Question answering systems, in contrast, deliver exact answers.

In order to correctly answer a question, one first usually needs to understand what the question asks for. For instance, if a QA system understands that the question *Hver er höfuðborg Hollands?* (e What is the capital of the Netherlands) has an answer type of city, extracting a correct answer becomes more feasible; the answer space is decreased by orders of magnitude. For this reason, the accuracy of question classification can heavily affect the overall performance of a question answering system.

Although question classification can be performed using hand-written rules, such an approach requires deep linguistic knowledge of the language. In order to get such a system achieving high accuracy, a great deal of different kinds of questions must be taking into account. It might seems ideal, for instance, to classify a question starting with *Hvaða ár...* (e. Which year) to have an answer type of year. However, one would later realize that *ár* can also be the plural noun for *rivers* and such a question (e.g *Hvaða ár renna saman í Miðfirði?*, e which rivers merge in Miðfjörður fjord) would then have the answer type of location. The only QA system developed for the Icelandic language, so far, relied on hand written regular expressions as these mentioned above to perform question classification. The outcome of that approach was that the system could only confidently answer three types of questions, i.e. those having an answer type of year, person or location (Geirsson, 2013). Such a compromise defeats the objective of developing an open-domain QA system.

An alternative option is to employ machine learning techniques. Such an approach has been taken by state-of-the-art systems resulting in up to a 90% classification accuracy (Huang et al., 2008; Zhang and Lee, 2003). Though such systems, being data-driven, require large amounts of annotated text to work, they eliminate the need for deep linguistic intuition. Furthermore, existing software to build a classification model allows one to rapidly test this approach.

This report describes such an attempt with the Icelandic language. In Section 2 we discuss the preparation of a new Icelandic corpus for use in question classification, in Section 3 we look at an experimental study on how a question classifier performs on this corpus and in Section 4 we make our conclusions.

2 Building a corpus

In order to build a question classification model, one must first have hold of annotated questions. For this research, a large set of question we're collected and then annotated by their expected answer type. A commonly used answer type taxonomy was adopted.

2.1 Taxonomy

The Li and Roth (2002) answer type taxonomy is widely used in question classification research (Huang et al., 2008; Zhang and Lee, 2003). Li and Roth also published the UIUC[1] corpus, a collection of over 5000 questions annotated using their taxonomy. This corpus is often used as a benchmark for question classification accuracy. The taxonomy is two-layered and consists of 6 coarse classes and 50 fine classes. The full taxonomy is listed in Table 2.1

Some classes in the taxonomy seemed to conflict, mainly in a way involving word-sense disambiguation. An institution, for example, can fall into any of the categories *Human:Group*, *Entity:Other* (as in institution) and *Location:Other* (as in building). Such uncertainties were resolved by looking up similar cases in the UIUC corpus. In the case of no clear solution, the annotator was left to decide himself which label felt most appropriate.

During annotation, a need to introduce a new coarse category came up, namely the *Yes/No* class. The answer to a yes/no question is either one of two available options, e.g. *Er Spánn sunnar en Japan?* (e. Is Spain further south than Japan?). Event though Li and Roth taxonomy does not allow such a category, one can imagine it being useful in the context of open-domain QA systems.

Coarse classes	Fine classes
Abbreviation	abb, exp
Description	definition, description, manner, reason
Entity	animal, body, color, creative, currency, dis.med., event, food, instrument, lang, letter, other, plant, product, religion, sport, substance, symbol, technique, term, vehicle, word
Human	group, individual, title, description
Location	city, country, mountain, other, state
Numeric	code, count, date, distance, money, order, other, period, percent, speed, temp, size, weight

Table 1: Li and Roth answer type taxonomy

Additionally, a lack of certain classes was found while annotating questions. For instance, the class *Entity:creative* covers all movies, songs, books, poems and other creative entities. Considering the popularity of such questions (e.g. in quiz shows) one might be tempted to separate them. However, to keep consistency with the UIUC corpus such measures were not taken. Instead, comments were added to the side which might turn out to be useful later on.

2.2 Gettu Betur corpus

Gettu Betur (e. Make a Better Guess) is a popular Icelandic team quiz show amongst high school students, broadcast on the national television channel RÚV. It is an open-domain quiz show where knowledge is tested on well known figures and facts about history, geography, literature, natural sciences and more.

A collection of over 20.000 Gettu Betur questions were donated from a local school team. This collection has been used since 2003 as revision material for the opening round of the show, i.e hraðaspurningar (e. quick questions). During *Hraðaspurningar* a series of factoid questions — questions whose answers are based on factual information — are asked under strict time limits. Given

[1] Available at http://cogcomp.cs.illinois.edu/Data/QA/QC/

	Gettu Betur		UIUC	
Fine classes	#	%	#	%
Abbreviation	16	0.35%	86	1.58%
Description	137	**3.00%**	1162	**21.31%**
Entity	711	15.55%	1250	22.93%
Human	2108	**46.11%**	1223	**22.43%**
Location	1112	24.32%	835	15.32%
Numeric	467	10.21%	896	16.43%
Yes/no	21	0.46%	0	0.00%
Sum	4572		5452	

Table 2: Comparison of coarse category distributions for Gettu Betur and UIUC corpora. Boldface: Clear difference between the two corpora

the time constraints, answers are often relatively short and rarely require a lengthy explanation or enumerations of long list.

A random subset of 4569 questions from the collection were annotated, referred to as the *Gettu Betur* corpus from now. The distribution of coarse category questions in the Gettu Betur and UIUC corpora can be found in Table 2.2. An observation on this table is worth pointing out; the categories in the UIUC corpus spread more evenly. For instance, the human category dominates nearly half of the Gettu Betur questions while the description category is very small and nearly seven times larger in the UIUC corpus. One explanation for this might be that the UIUC corpus was specifically prepared for research and contains questions from a wide range of sources (Li and Roth, 2002) while the Gettu Betur corpus questions come from a single source.

3 Experimental study

We designed two experiments. The first experiment was to test which features might would be most useful training a classifier. For this part, we compare the accuracy of the classifier on both the UIUC and Gettu Betur corpora. The second experiment was to test how the size of the training set would affect the accuracy of the classifier. For this part, we only used the Gettu Betur corpus.

3.1 Training model

Maximum entropy models have shown to work well on text classification (Berger et al., 1996). In the following, the Stanford classifier (Manning and Klein, 2003) is used. It implements a maximum entropy classifier and comes with built in features useful, in particular, for text classification.

The features tested were as follows

Char n-gram Characters sequences, including spaces and punctuations. The number appearing in brackets in the tables denotes the maximum length of the sequences used.

Word n-gram Word sequences split on whitespace. A number in brackets denotes the maximum length of the sequences used. Includes *Char n-grams*.

Filter Filter out stop words, namely adverbs and infinitives. Includes *Char + Word n-grams*.

Lemma Lemmatize question in order to overcome the sparse data problem. Includes *Char + Word n-grams*.

POS Perform part of speech tagging. The tags are separated from the text into another column. Conceivably, the classifier might be able exploit the syntax of the question to identify a correct class. Includes *Char + Word n-grams*.

HWs *Headword* extraction. Headwords are the first words appearing in each question up to and including the first verb (e.g. *Who is* from “Who is the president of Iceland?”). Conceivably, a higher weight on headwords might assist the classifier in identifying a correct class. Includes *Char + Word n-grams*.

A flat (and naive) approach for performing fine category classification would be to let the classifier directly predict a label from one of the 50 fine classes. However, this would not make use of the two-layered taxonomy. Instead, a hierarchical classifier (Li and Roth, 2002) was used. With this approach, a category predicted by a coarse classifier was used as feature for the fine category classification. Unlike the classifier explained by Li and Roth however, the classifier in this experiment did not output multiple labels.

3.2 Data

A random subset of 4500 questions from the *Gettu Betur* and *UIUC* corpora were used in the following. In training, we divide the 4500 questions randomly into nine even chunks and then perform nine-fold cross validation.

3.3 Evaluation

The evaluation metric used for the general performance of the classifier is a simple accuracy score calculated as the fraction of correctly labeled questions by the classifier divided by the number of total questions. The total number of questions was obtained from summing the results over all nine runs during cross-validation.

For evaluating the performance of the classifier in classifying each specific class a combination of precision, recall and F1 measure were used. For each respective class, the total number of *true positives*, *true negatives* and *false negatives* are summed from all nine runs of cross-validation. The precision is then measured as

$$\text{precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}},$$

the recall is measured as

$$\text{recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

and the balanced $F1$ measure score calculated as

$$\text{F1} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

The precision gives a representation of the fraction of correctly selected questions from the total number of all selected questions. For instance, if the classifier identifies 50 out of 100 questions to have an answer type of *human* while only 20 of them are correct, the precision would be $\frac{20}{50} = 40\%$. The recall gives a representation of the fraction of correctly classified questions from the total number of questions in the given class. Following the previous example, say that there are a total of 30 questions in the test set having the answer type of *human* the recall would then be calculated as $\frac{20}{30} \approx 67\%$. The F1 measure is a balanced average of these two measurements.

Features	Gettu Betur	UIUC
Char unigram	60.09%	48.18%
Word unigram	88.49%	82.84%
Char four-gram	90.04%	83.42%
Char six-gram	91.40%	85.42%
Word four-gram	90.27%	84.76%
Filter	88.86%	—
Lemma	87.48%	—
PoS	89.64%	—
HWs	92.13%	—

Table 3: Accuracy of coarse category classifier. Training set size = 4.000 questions.

Features	Gettu Betur
Word four-gram	84.98%
HWs	85.58%

Table 4: Accuracy of fine category classifier. Training set size = 4.000 questions.

3.4 Experimental results

Table 3.4 shows the accuracy of the coarse classifier with respect to different features. The results are quite encouraging; the highest achieved accuracy is 92.13% using headwords. The comparison with the *UIUC* corpus, however, shows that the classifier does not perform as well on a more diverse set of questions. The character and word n-grams perform reasonably well considering they do not require any further text processing, in contrast with the headwords. The *filter*, *lemma* and *PoS* on the other hand, perform surprisingly worse.

Table 3.4 shows the accuracy of the fine classifier with respect to the two most promising features, word four-grams and headwords. As suspected, the accuracy is lower than for the coarse classifier. However, the drop is somewhat comparable to state-of-the-art systems (Zhang and Lee, 2003; Huang et al., 2008).

Figure 3.4 shows the relation between the training set size and accuracy for both the fine and coarse

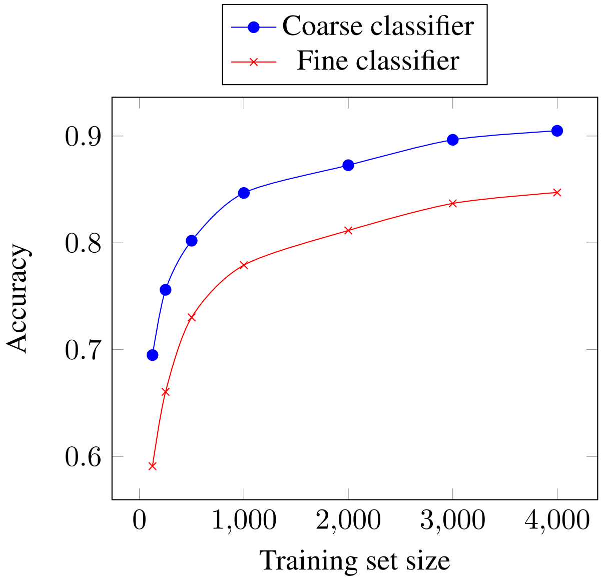

Figure 1: Performance improvements by increasing training set size

Coarse class	Precision	Recall	F1	#
Numeric	0.98	0.94	0.96	459
Human	0.93	0.97	0.95	2077
Location	0.95	0.93	0.94	1095
Entity	0.85	0.81	0.83	700
Description	0.84	0.74	0.79	134
Yes-No	0.93	0.65	0.76	20
Abbreviation	0.90	0.60	0.72	15

Table 5: Precision, Recall, and F1 measure for coarse category classification with Gettu Betur corpus using four-gram character, four-gram words and headwords.

classifiers. The coarse classifier performs surprisingly well with a train set 2.000 questions by achieving almost 85% accuracy; doubling the training set size only yield a roughly 3% increase in accuracy. The improvement from increasing the train set of 3.000 questions to 4.000 is less than one percent, telling us the accuracy would quickly converge to approximately 91% with larger training sets. A similar trend is repeated for the fine classifier except that it converges to approximately 85% accuracy.

Table 3.4 shows the *precision*, *recall* and *F1* measure of the coarse classifier in predicting each of the coarse classes. The first observation is that the overall precision of the classifier is higher than the overall recall; this tells us the classifier favors larger classes to smaller classes. The second observation is that the classifier performs, unsurprisingly, well on the larger classes while it performs a lot worse on the smaller classes.

Table 3.4 shows the same measurements as Table 3.4 except for the fine classifier in predicting each one of the finer classes. Similar trends can also be found here; the overall precision is again higher than the overall recall and the larger the classes are the more accurately the classifier is able to predict them. A closer observation also tells us that the classifier generally struggles more with the *other* classes than the predefined classes. This is plausible since questions in *other* classes don't necessarily share a common structure.

3.5 Discussion and examples

We have shown that the overall accuracy of our classifier is satisfactory. Indeed, it performs not far from state-of-the-art classifiers for the English language. Nevertheless, it is constructive to consider some cases in which the classifier fails. Below are some examples misclassified by the fine classifier.

Hver voru hin sovésku tákn verkamanna og bænda?
(e. What were the symbols for Soviet workers and farmers?)

The correct label is *Entity:Symbol* while the classifier — probably mislead by the beginning "hver" (e. who) — outputs *Human:Individual*. The word "tákn" (e. symbol), however, gives us the correct answer so it might be possible for the classifier to accurately make a prediction if given more training examples.

Hve mörg kíló efnis umbreytast í orku í 20 Megatonna kjarnorkusprengju?
(e. How many kilos of material are transformed into energy in a 20 megaton nuclear bomb?)

The correct label here is *Numeric:Weight* while the classifier outputs *Numeric:Count*. However, given that the questions is phrased this way one could argue that this output is not strictly incorrect.

Hvar á landinu var Milljónafélagið starfrækt?
(e. Where in the country did the *Milljónafélag* operate?)

Fine class	P	R	F1	#
Numeric:Date	0.95	0.98	0.97	315
Entity:Color	1.00	0.93	0.97	30
Human:Individual	0.90	0.98	0.94	1944
Location:City	0.93	0.89	0.91	389
Location:State	1.00	0.83	0.91	29
Numeric:Code	1.00	0.81	0.89	21
Numeric:Count	0.88	0.91	0.89	74
Entity:Substance	0.87	0.90	0.89	52
Location:Country	0.90	0.85	0.88	278
Entity:Lang	0.92	0.83	0.87	29
Abbreviation:Abb.	1.00	0.75	0.86	12
Entity:Sport	1.00	0.72	0.84	18
Yes-No	1.00	0.71	0.83	21
Entity:Symbol	1.00	0.67	0.80	9
Location:Other	0.74	0.79	0.77	373
Numeric:Period	1.00	0.62	0.77	29
Description:Def.	0.63	0.95	0.76	20
Description:Des.	0.75	0.71	0.73	87
Entity:Creative	0.66	0.75	0.71	158
Entity:Animal	0.74	0.68	0.71	62
Entity:Term	0.58	0.85	0.69	102
Location:Mount.	0.92	0.52	0.67	21
Description:Reason	0.89	0.50	0.64	16
Human:Group	0.75	0.50	0.60	124
Entity:Event	1.00	0.29	0.44	21
Human:Title	1.00	0.27	0.42	15
Entity:Other	0.43	0.39	0.41	101
Entity:Religion	1.00	0.17	0.29	6
Entity:Body	1.00	0.12	0.21	17
Description:Manner	1.00	0.09	0.17	11
Abbreviation:Exp	0.00	0.00	0.00	4
Entity:Dis.Med	0.00	0.00	0.00	5
Entity:Instrument	0.00	0.00	0.00	3
Entity:Letter	0.00	0.00	0.00	3
Entity:Plant	0.00	0.00	0.00	12
Entity:Product	0.00	0.00	0.00	4
Entity:Vehicle	0.00	0.00	0.00	16
Entity:Word	0.00	0.00	0.00	14
Numeric:Distance	0.00	0.00	0.00	7
Numeric:Order	0.00	0.00	0.00	5
Numeric:Other	0.00	0.00	0.00	4

Table 6: Precision P, Recall R and F1 measure for fine category classification with Gettu Betur corpus using four-gram character, four-gram words and headwords. Classes containing no labeled questions are left out. Boldface: Fine categories labelled as "Other" perform on average worse

The correct label here is *Location:Other* while the classifier outputs *Location:City*. The annotator in this case did not assume that the company would necessarily have to operate in a city and therefore labeled the question as *other*. As with the previous example however, one could argue that this output is not strictly incorrect.

Finally, it is interesting to mention that the classifier was able to identify a couple of questions that were mislabeled by the annotator. We found this out by sorting the wrongly predicted questions after the confidence score output from the classifier.

4 Conclusions

This report presents a machine learning approach to question classification for the Icelandic language. We adopted a commonly used hierarchical taxonomy and we built a corpus, *Gettu Betur*, of annotated questions in Icelandic on which we trained a maximum entropy classifier using six different features: character n-grams, word n-grams, filtering stopwords, lemmatization, part of speech tagging and headwords, i.e. filtering out words following the first occurring verb. Character and word n-grams proved to perform well while features requiring further text processing generally lead to decreased performance. The only feature resulting in a higher accuracy was headword extraction.

Questions in the new *Gettu Betur* corpus do not spread out as evenly across the taxonomy as the more commonly used *UIUC* corpus. Our experimental results show that this learning approach is very promising and achieves up to 92% classification accuracy on coarse classes and 85% on fine classes, a higher score than achieved by state-of-the art systems. However, our experiment also shows that the classifier performs better on questions in classes which appear more often in the corpus. We suspect this fact to be the reason why the classifier performs so well.

In future work we plan to investigate further both how the classifier performs on a more diverse set of questions and also how the accuracy is impacted by allowing the classifier to output multiple labels. Moreover, we plan to release a trained classifier for use in an open-domain question answering research for the Icelandic language.

References

[Berger et al.1996] Adam L. Berger, Vincent J. Della Pietra, and Stephen A. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Comput. Linguist.*, 22(1):39–71, March.

[Geirsson2013] O. Geirsson. 2013. IceQA: A Question Answering System for the Icelandic language. Technical report, Reykjavik University, School of Computer Science, 8.

[Huang et al.2008] Zhiheng Huang, Marcus Thint, and Zengchang Qin. 2008. Question classification using head words and their hypernyms. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, page 927–936, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Li and Roth2002] Xin Li and Dan Roth. 2002. Learning question classifiers. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, page 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Manning and Klein2003] Christopher Manning and Dan Klein. 2003. Optimization, maxent models, and conditional estimation without magic. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Tutorials - Volume 5*, NAACL-Tutorials '03, pages 8–8, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Zhang and Lee2003] Dell Zhang and Wee Sun Lee. 2003. Question classification using support vector machines. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, SIGIR '03, page 26–32, New York, NY, USA. ACM.