Abstract

Question Answering (QA) systems are the next generation of search technology. A QA system exploits more information about a user’s intention, returning the exact answer to a users query. A crucial component of a QA system is Question Classification (QC), which involves determining the type of information the question is requesting. QC narrows the scope of processing within a QA. For example, “Who is the fastest swimmer?” - the expected answer type (or classification) could be HUM:ind, which denotes a human individual. This project develops a new approach to QC, which involves the application of statistical methods, namely Maximum Entropy models for classification. This approach outperforms state-of-the-art accuracy.

1 Introduction

Question Classification is an important component of QA systems. QC is necessary as it identifies the answer type for a given question. This information places a constraint on the answer type. Various methods have been applied to QC. The majority of current QC systems make use of Machine Learning techniques. It is difficult to compare existing QC systems as they vary significantly in the answer type typology used.

Maximum Entropy models are an emerging technique in the field of NLP. These models have been successfully applied to many areas such as POS tagging and named entity recognition. The goal for this project is to classify questions with respect to their answer types using Maximum Entropy models.

2 Question Answering Systems

Burger et al. (2002) presents a road map for Question Answering and Text Summarization research. System architecture guidelines are identified and issues are explored. Components recommended include: question classes, question processing, context and QA, data sources, answer extraction and answer formation. A diagram reflecting the ideas of Burger et al. (2002), can be seen in Figure 1. All current QA systems implement this structure to varying degrees of sophistication.

Figure 1 shows the interaction of the QC within a QA system. QC is used within the question processing and answer formation components.

3 Question Classification

Question Classification (QC) is a vital component of a QA system. Research on QA suggests the importance of classifying questions with respect to their answer type. This knowledge reduces processing and provides a feasible way to select correct answers from among the possible answer candidates. More specifically, the entire passage retrieval and answer extraction process relies on having the correct answer type for a given question.

The current state-of-the-art in QC (Li and Roth, 2002), uses Machine Learning techniques. They make use of the Sparse Network of Winnows (SNoW) classifier. An accuracy of 84.2% is achieved on fine-grained classes and 91% on coarse-grained classes.

4 Project Objectives

Current Question Classification systems utilise Machine Learning (ML) techniques. The goal for this project is to classify questions with respect to their answer types. The project involves exploiting the advantages of Maximum Entropy models and apply it to QC, which to date, has not been attempted. The QC will be trained on the Li and Roth (2002) training data to determine its baseline performance and its final accuracy will be tested on the TREC-10 test set. The Li and Roth (2002) answer typology is used.

5 Features used in Experimentation

A variety of features were used during the experimental phase. The aim is to find the best combination of features to maximise
accuracy. Lexical features represent the words of a question as a feature. These include: raw word, lower case word - bringing all words down to lower case, stemmed word - using the Porter Stemmer to remove the suffix of a word and lemmatized word, which is similar to the Porter Stemmer however is a more sophisticated and accurate method. N-grams use multiple words within the feature. Bigrams represent all combinations of two consecutive words in a question. Trigrams represent all combinations of three consecutive words in a question. Additionally, the first two words of the question are represented as a feature. Syntactical features represent the syntactical aspects of a question. This includes: part of speech (POS) tags, which represent the grammatical functions of the question, chunks represent non-recursive grammatical phrase structure. Syntactic dependencies are also represented, which are produced by the Combinatory Categorical Grammar (CCG) parser. These dependencies describe the relationships between words in a question. The length of the question is also represented which denotes the number of words in the question. Semantic features are also represented. The target represents the focus of the question. It signifies what the question is actually asking. An algorithm was created, which uses both the syntactic dependencies produced by the CCG parser and POS information. Named Entities are also determined. Words, which are recognized as Named Entities (Person, Organization, Location) are represented as features. Lists of semantically related words (SRW), produced by Li & Roth (2002) are made publically available and have been used to group related words together.

Special expansion features utilise WordNet (Fellbaum, 1998), which is used to extract hypernyms for each noun in the question and synonyms for each verb in the question.

6 Naive Bayes

Naive Bayes is an example of a statistical method, which has been used successfully for various classification NLP tasks. Despite the poor assumption of independence, in practice, Naive Bayes performance is comparable to more sophisticated classification methods. A Naive Bayes classifier has been developed to provide a baseline for performance against Maximum Entropy.

Table 1 shows a sample of results for the Naive Bayes QC.

<table>
<thead>
<tr>
<th>Features</th>
<th>Fine-grained Result</th>
<th>Coarse-grained Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. 1st 2 words, target, length</td>
<td>73.4%</td>
<td>76.4%</td>
</tr>
<tr>
<td>2. 1st 2 words, target, len, POS</td>
<td>59.6%</td>
<td>79.2%</td>
</tr>
<tr>
<td>3. bigrams</td>
<td>28.8%</td>
<td>67.4%</td>
</tr>
<tr>
<td>4. all features</td>
<td>36.6%</td>
<td>68.8%</td>
</tr>
</tbody>
</table>

Table 1: Naive Bayes QC Results

It can be seen that the best combination of features for fine grained classification is the first two words of a question, its target or focus and the question length. An accuracy of 73.4% is competitive with current systems. The addition of POS tags allowed for an increase in coarse-grained accuracy to 79.2%. The addition however, hindered the fine-grained accuracy as it added a large amount of confusion to the classifier. The experiment using all features highlights that the task of QC is about selecting the best possible combination of features.

7 Maximum Entropy

Maximum Entropy (MaxEnt) models allow us to integrate information from many heterogenous information sources for classification. The features chosen are used as constraints on the model. A model is selected, which satisfies the constraints imposed by the features. Ideally, this enables the specification of all potentially relevant information at the beginning, and then to let the training procedure determine about how to come up with the best model for classification.

The over-riding principle in MaxEnt is that when nothing is known, the distribution should be as uniform as possible i.e. have maximum entropy.

8 Results

Table 2 shows a sample of results for the Maximum Entropy classifier.

The Maximum Entropy approach produces a state-of-the-art in Question Classification. The second experiment has yielded a result, which is 1.2% above the world standard. Utilizing the best combination of features, the classifier produced a model which classifies with an accuracy of 85.4%. Experiment 3 shows the features required to obtain a result equivalent to that of Li and Roth (2002). The clear advantage of Maximum Entropy over Naive Bayes is that it does not assume conditional independence between features.

9 Conclusion

This project was motivated by the need to increase the accuracy of Question Answering systems. To do this a technique was applied that has not yet been explored. By improving on the quality and accuracy of Question Classification, QA systems can more reliably predict the answers to questions posed by users. Maximum Entropy has achieved a state-of-the-art accuracy of 85.4%.

References

