Extracting the Unknown from Long Math Problems

Ndapa Nakashole
Computer Science and Engineering
University of California, San Diego
La Jolla, CA 92093
{nnakashole}@eng.ucsd.edu

Abstract

In problem solving, understanding the problem that one seeks to solve is an essential initial step. In this paper, we propose computational methods for facilitating problem understanding through the task of recognizing the unknown in specifications of long Math problems. We focus on the topic of Probability. Our experimental results show that learning models yield strong results on the task — a promising first step towards human interpretable, modular approaches to understanding long Math problems.

1 Introduction

"It is foolish to answer a question that you do not understand. It is sad to work for an end that you do not desire" — George Polya (Polya, 1957). According to Mathematician Polya’s famous problem solving heuristics, asking general common sense questions about the problem facilitates problem understanding. One such question is “What is the unknown?” (Polya, 1957). The unknown of a given problem is what the problem requires to be worked out and solved. In this paper, we propose computational models for extracting the unknown from problem specifications.

In online tutorial communities, it is not uncommon to find problems that include questions like “what is the solution I am looking for?”1. Automated hints such as help with identifying the unknown, could provide timely support to self-directed learners (Guglielmino, 1978; Houle, 1988; Hiemstra, 1994). Furthermore, information about the unknown can be used as a feature in other tasks that can enhance self-directed learning, such as similar-question retrieval (Lei et al., 2016), and question-answer matching (Shen et al., 2017), as well as in models for automated problem solving (Sachan et al., 2018).

The unknown is general, it not restricted to problems about a particular subject-matter. The problem can be algebraic, geometric, mathematical or nonmathematical (Polya, 1957). In this paper, we focus on the topic of Probability as a case study. To facilitate model training, we created a dataset, PROBUNK, containing Probability problems labeled with the unknown(s). Figure 1 shows example problems from PROBUNK.

Prior Work. Our work is related to question understanding, and modular approaches to question answering that advocate decomposing questions into smaller ones (Andreas et al., 2016; Iyyer et al., 2017; Gupta and Lewis, 2018; Talmor and Berant, 2018; Huang et al., 2018; Guo et al., 2019; Min et al., 2019; Wolfson et al., 2020). However, question decomposition in prior work is aimed at en-

Figure 1: Example problems from PROBUNK, on the concepts of top: Correlation, bottom: Normal Distribution, labeled with the unknown.
hancing model performance, and is generally not human friendly. Notions of interpretability are provided through attention modules (Andreas et al., 2016), or explicit meaning representations such as first-order logic (Sachan et al., 2018), or custom semantic parsing formalisms (Amini et al., 2019; Wolfson et al., 2020). In contrast, our goal is to produce output that can be used by humans, and potentially by models too. Additionally, questions in most prior work are short, containing just a single sentence. We target longer multi-sentence problems.

2 Dataset

Background Knowledge on Probability. As background knowledge, we collected information about probability concepts. We define a probability concept as a term that is formally defined in the first five chapters of Wasserman (2013), and appears in the index at the end of that book. This produced 69 concepts. For example, ‘Poisson Distribution’, ‘Correlation’, see appendix for full list. We augment each concept with content from DeGroot and Schervish (2012). On average per concept, we collected 1.6 definitions 1.04 worked-out examples, and 1.54 unworked-out examples, from both books.

StackExchange Probability Problems. We obtained the data dump from stats.stackexchange, containing 281, 265 Statistics problems. Each problem is labeled with relevant concept tags. We discarded all programming-related problems, tagged with: ‘Matlab’ and ‘R’, resulting in 139, 303 problems. We also pruned concept tags outside of the 69 background concepts. To avoid overly complex problems, those with > 3 tags were pruned. Lastly, we only kept problems for which there is an ‘accepted answer’ from the forum. Overall after pre-processing, there were 1,171 remaining problems, spanning 11 Probability concepts, along with their answers. Table 1 shows the train/dev/test split.

| Method | Dev F1 | Test F1 | Training time |
|--------|--------|---------|---------------|
| 1. MaxEnt | 0.534 | 0.673 | 04m32s |
| 2. MLP | 0.547 | 0.636 | 06m24s |
| 3. LSTM | 0.519 | 0.561 | 28m41s |
| 4. GRU | 0.537 | 0.567 | 25m36s |
| 5. CNN | 0.754 | 0.727 | 03m39s |
| 6. Prototypical Nets(*) | 0.787 | 0.782 | 03m21s |

Table 2: Multi-label concept classification.

Figure 2: Concept prototypes (labeled circles) and prototypical examples (diamonds).

Table 1: Dataset statistics and split.

| Problems | Dev | Test |
|----------|-----|------|
| 904      | 110 | 157  |
| Answers  | 904 | 110  | 157  |
| Total    | 1,808 | 220  | 314  |

Table 2: Multi-label concept classification.

2Breakdown of problems by concept is in the appendix.
concepts. This is a multi-label classification task. We used a pretrained transformer, BERT (Devlin et al., 2019), to obtain contextualised word embeddings, and then trained various models. The results are shown in Table 2. For the MaxEnt model (a logistic regression classifier), and the multi-layer perceptron, MLP, problem representations were obtained by average pooling of the word embeddings. From initial experiments we found that using the CLS token representation, instead of average pooling, degrades performance. A Convolutional Neural Network (CNN) text classifier, kernel sizes: 3, 4, 5, 6, each with 192 kernels, has the strongest performance on the full problem, F1 0.73 on test data. We also implemented Prototypical Networks (Snell et al., 2017), an approach designed for limited labeled data settings. We used a CNN (same settings as above) as the base model for learning prototypical vector representations. We evaluated the Prototypical Network on problems that are about a single concept, 72.7% of dev, and 75.8% of test data, and it produced strong results, F1 0.78. Visualizing the learned prototypes and prototypical problems, Figure 2 shows that the learned representations are meaningful. For instance, prototypical problems of type Poisson Distribution are closest to the prototypical vector of that concept. This is not trivially expected because there are many prototypical vectors for a concept, which are generated from different support vectors (refer to Snell et al. (2017) for details), and we randomly picked one prototypical vector for each concept.

Overall, both the quantitative results in Table 2, and the qualitative results in Figure 2, show that our dataset is adequate for learning meaningful representations, when using pre-trained contextual embeddings.

Unknown Annotation. To collect labeled data, we assume the unknown is a contiguous sequence of tokens in the problem specification. We implemented a simple labeling tool, shown in Figure 3. A problem can have multiple unknowns, each unknown is entered separately. In total, 1,171 problems were annotated. Problems whose unknowns were deemed ‘unclear’ were few: 26/904(2.9%), 4/110(3.6%), and 2/157(1.3%) for train, dev, test,
respectively. On average the unknown(s) is (are) spread across 1.7. Figures 4(a) and 4(b) show distributions of sentences per problem, and position of the unknown.

3 Approach

Given a sentence, our goal is to predict if it contains a sequence of words that describe the unknown, $y_u = 1$, or not, $y_u = 0$. Alternative formulation is to treat the problem a sequence labeling task. However, in this initial work, the labeled data we obtained is at the sentence level. With the right granularity of labeled data, our task be cast as a sequence labeling task.

Input, Prediction, and Loss. As input, a problem specification, $Q_i$, and the $j$th sentence, $s_{i,j} \in Q_i$, are presented to a model. A context vector, $c_i$, is generated from $Q_i$, and a sentence vector, $x_{i,j}$ is generated from $s_{i,j}$. The two vectors are concatenated to compute $p_u$ as follows:

$$p_u = \sigma(w^T[c_i; x_{i,j}] + b),$$

where $\sigma$ is the sigmoid function, $w$ and $b$ are learned via the cross entropy loss

$$\sum(y_u \log p_u).$$

Context Vectors. To compute the context vector, $c_i$, we consider different approaches.

BoW context: $c_i$ is obtained by average pooling BERT embeddings of the tokens of $Q_i$.

CNN Context: $c_i$ is generated by applying a CNN to $Q_i$.

GCN Context: $c_i$ is generated from a Graph Convolutional Network (GCN) (Bruna et al., 2014; Defferrard et al.; Kipf and Welling, 2017). Given a graph with $n$ nodes, the input to a GCN is a feature matrix $X \in \mathbb{R}^{n \times m}$ where a row contains the initial representation, $x_{u} \in \mathbb{R}^{m}$ of a node $u$. To compute a node’s hidden representation, $h_{u} \in \mathbb{R}^{d}$, a single graph convolution layer uses the node’s neighbors $N(u)$:

$$h_{u} = f \left( \sum_{v \in N(u)} (w_{u,v} + b) \right)$$

$f$ is a non-linear activation function, $w \in \mathbb{R}^{d \times m}$ and $b \in \mathbb{R}^{d}$ are learned. We apply the GCN recurrence for $K = 3$ steps. At every step, node representations are re-computed using the representation of the node’s neighbors from the previous step, see Kipf and Welling (2017) for details.

The goal is to use the graph to learn the kinds of unknowns that pertain to different concepts, as well as to exploit relationships between problems and answers. We construct a graph with three types of nodes: concept, problem; and answer nodes. We have 5 edge types: problem-has-type (between problems and concepts); problems-has-answer (between problems and answers); and same-section-as, mentioned-in-before-chapters, same-chapter-as (all three between concepts). We use the concept hierarchy derived from the ordering of concepts in Wasserman (2013) to obtain the last three relationships. Our graph consists of 2,360 nodes, 5,116 edges, and 768 initial node features obtained from BERT embeddings. Node features are from average pooling the embeddings. For the concept nodes, the pooled tokens are taken from the concept definitions. The task for learning node representations is predicting the problem-has-type links.

4 Experimental Evaluation

We conducted experiments to evaluate performance of various models on the task of extracting the unknown. All experiments are based on the train/dev/test split in Table 1. Training and computing infrastructure details are in the appendix.

Methods under Comparison. Next to each learning method is the number of parameters. 1) Majority. Assigns the most common label, $p_u = 0$ to every sentence. 2-4) $n$-th Sentence assigns $p_u = 1$ to the $n$-th sentence, and $p_u = 0$ to all others. 5) MaxEn (1, 537). Uses BoW for the context vector, $c_i$, and the sentence vector $x_{i,j}$. 6) MLP (1, 312, 769). Uses the same $c_i$ and $x_{i,j}$ as method (5). 7) CNN (1, 414, 275). A CNN generates both $c_i$, and $x_{i,j}$. 8) CNN_NoContext (1, 166, 467). Same as (7) but $c_i$ is omitted, thus the CNN only encodes the sentence. 9) CNN+Graph (1, 502, 386). Concatenates CNN and GCN contexts to get $c_i$, whereas $x_{i,j}$ is obtained from just the CNN. 10) CNN+Graph+LSTM (2, 980, 018). $c_i$ is as in (9), and $x_{i,j}$ is generated by an LSTM. 11) CNN+Graph+GRU (2, 610, 610). Same as (10) but $x_{i,j}$ is generated by a GRU.

Results. Table 3 shows our main results. Among non learning baselines, the 1st sentence (2) base-
Table 3: Unknown extraction F1.

| Method               | DEV F1 | TEST F1 | Training time |
|----------------------|--------|---------|---------------|
| 1. Majority          | 0.456  | 0.456   | n/a           |
| 2. 1st Sentence      | 0.657  | 0.657   | n/a           |
| 3. 2nd Sentence      | 0.499  | 0.528   | n/a           |
| 4. Last Sentence     | 0.531  | 0.508   | n/a           |
| 5. MaxEnt            | 0.670  | 0.686   | 02m19s        |
| 6. MLP               | 0.733  | 0.723   | 03m06s        |
| 7. CNN               | 0.802  | 0.776   | 10m31s        |
| 8. CNN_NoContext     | 0.777  | 0.759   | 09m10s        |
| 9. CNN + Graph       | 0.802  | 0.780   | 16m42s        |
| 10. CNN + Graph + LSTM | 0.504  | 0.499   | 19m48s        |
| 11. CNN + Graph + GRU | 0.495  | 0.492   | 18m44s        |

What cues are the methods relying on to achieve strong performance on this task? We sampled some sentences from data, a few of which are shown in Table 4. It is clear that unambiguous cues for our task are indeed in the sentences, such as “derive”, “calculate”, “what is the probability that ”, and “prove that ”. In addition to these sentence cues, our results, (7) vs. (8) and (9) in Table 4 show that there are also useful signals in the context.

5 Conclusions and Future Work

Our new task, dataset, and results are only initial steps, and pave the way for future work. Beyond “what is the unknown?”, we can develop models that ask other general common sense questions that can support the problem solving process. Polya (1957) presents a list of more such questions. These questions are general, future work includes working on problems from topics other than Probability.
Additionally, future work can study who this line of work can benefit from well-studied aspects of knowledge harvesting and representation (Nakashole et al., 2010; Theobald et al., 2010; Nakashole et al., 2012; Nakashole, 2012; Kumar et al., 2017).

References

Aida Amini, Saadia Gabriel, Peter Lin, Rik Koncel-Kedzioraki, Yejin Choi, and Hannaneh Hajishirzi. 2019. Mathqa: Towards interpretable math word problem solving with operation-based formalisms. arXiv preprint arXiv:1905.13319.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Learning to compose neural networks for question answering. In NAACL HLT, pages 1545–1554. The Association for Computational Linguistics.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. 2014. Spectral networks and locally connected networks on graphs. In ICLR.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. Moriss H DeGroot and Mark J Schervish. 2012. Probability and statistics. Pearson Education.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In NAACL-HLT, pages 4171–4186.

Lucy M Guglielmino. 1978. Development of the self-directed learning readiness scale. Ph.D. thesis, ProQuest Information & Learning.

Jiaqi Guo, Zecheng Zhan, Yan Gao, Yan Xiao, Jian-Guang Lou, Ting Liu, and Dongmei Zhang. 2019. Towards complex text-to-sql in cross-domain database with intermediate representation. In ACL, pages 4524–4535.

Nitish Gupta and Mike Lewis. 2018. Neural compositional denotational semantics for question answering. In EMNLP, pages 2152–2161.

R Hiemstra. 1994. Self-directed learning. In T. Husen & T N Postlethwaite (Ed.). The International Encyclopedia of Education (2nd). Oxford: Pergamon Press.

Elaine K Horwitz, Michael B Horwitz, and Joann Cope. 1986. Foreign language classroom anxiety. The Modern language journal, 70(2):125–132.

Cyril O Houle. 1988. The inquiring mind: A study of the adult who continues to learn.

Danqing Huang, Jin-Ge Yao, Chin-Yew Lin, Qingyu Zhou, and Jian Yin. 2018. Using intermediate representations to solve math word problems. In ACL, pages 419–428.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In ACL, pages 1821–1831.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In ICLR.

Anurag Kumar, Bhiksha Raj, and Ndapandula Nakashole. 2017. Discovering sound concepts and acoustic relations in text. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, pages 631–635. IEEE.

Tao Lei, Hrishikesh Joshi, Regina Barzilay, Tommi S. Jaakkola, Kateryna Tymoshenko, Alessandro Moschitti, and Lluis Marquez. 2016. Semi-supervised question retrieval with gated convolutions. In NAACL HLT, pages 1279–1289.

Sewon Min, Victor Zhong, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019. Multi-hop reading comprehension through question decomposition and rescoring. In ACL, pages 6097–6109.

Ndapandula Nakashole, Mauro Sozio, Fabian M. Suchanek, and Martin Theobald. 2012. Query-time reasoning in uncertain RDF knowledge bases with soft and hard rules. In Proceedings of the Second International Workshop on Searching and Integrating New Web Data Sources, Istanbul, Turkey, August 31, 2012, volume 884, pages 15–20.

Ndapandula Nakashole, Martin Theobald, and Gerhard Weikum. 2010. Find your advisor: Robust knowledge gathering from the web. In Proceedings of the 13th International Workshop on the Web and Databases 2010, WebDB 2010, Indianapolis, Indiana, USA, June 6, 2010. ACM.

Ndapandula T Nakashole. 2012. Automatic extraction of facts, relations, and entities for web-scale knowledge base population.

Mark Palatucci, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell. 2009. Zero-shot learning with semantic output codes. In NIPS, pages 1410–1418.

George Polya. 1957. How to solve it. Doubleday, Garden City, NY.

Bo Qu, Gao Cong, Cuiying Li, Aixin Sun, and Hong Chen. 2012. An evaluation of classification models for question topic categorization. J. Assoc. Inf. Sci. Technol., 63(5):889–903.

Manmaya Sachan, Kumar Avinava Dubey, Tom M. Mitchell, Dan Roth, and Eric P. Xing. 2018. Learning pipelines with limited data and domain knowledge: A study in parsing physics problems. In NeurIPS, pages 140–151.
6 Dataset

6.1 Prototypes Visualization Further Details
We defined prototypical examples as those whose closest prototypical vector is almost always (in ≥ 95% of the cases) that of the gold truth concept, regardless of the choice of support problems used to generate the prototypical vector. We then randomly picked one of the prototypical vectors, plotted it, and a few of the prototypical example problems.

6.2 Background Knowledge Concepts
The 69 Background knowledge concepts collected in our dataset shown in Table 5. Definitions, along with worked out example for each concept are in the dataset.

6.3 Stats.StackExchange Preprocessing
Excluded and ignored tags are shown in Table 7. If the tag is excluded, we remove from our dataset problems with this tag. If a tag is ignored, we can keep the problem but only if it has other tags besides an ignored tags.

6.4 Train/Dev/Test by Concept
Train/Dev/Test split of our dataset listing the number of problems in each partition per concept is shown in Table 6.

6.5 Concept Classification Training Details

Implementation. In our implementation, we used the Pytorch Library.

Computing Infrastructure. OS: 16.04.1-Ubuntu/ x86_64.
GPU: GeForce GTX 1080 Ti with 12 GB of memory.

Evaluation Metrics. Evaluation metric is Macro f1_score: We used the implementation in sklearn. In python, the code is: 'from sklearn.metrics import f1_score'

Training and Method Details. We use the ADAM optimizer, and dropout 0. Word embeddings are initialized with pretrained 768 dimensional BERT embeddings. The MLP has 3 hidden layers, and the dimensionality of hidden states is 512. The LSTM is a one layer bidirectional LSTM, with hidden dimensions 384. Concatenating backward and forward outputs of BiLSTM generates LSTM output vector of size 768. The GRU is a one layer bidirectional GRU, with hidden dimensions of 384. Concatenating backward and forward outputs of BiGRU generates GRU output vector of size 768.

Table 8 shows performance on the dev set over 11 seeds are used.

Parameter details for the methods for concept classification are in Table 9.
| Concept                                      | #. Items |
|---------------------------------------------|----------|
| **Probability**                             |          |
| 1. Probability                              | 2        |
| 2. Complement Of An Event                   | 2        |
| 3. Disjoint                                 | 3        |
| 4. Event                                    | 12       |
| 5. Intersection Of Events                   | 2        |
| 6. Monotone Increasing                      | 2        |
| 7. Mutually Exclusive                       | 3        |
| 8. Partition                                | 2        |
| 9. Sample Outcome                           | 2        |
| 10. Sample Space                            | 12       |
| 11. Set Difference                          | 1        |
| 12. Union                                   | 1        |
| 13. Uniform Probability Distribution        | 7        |
| 14. Independent Events                      | 11       |
| 15. Conditional Probability                 | 13       |
| 16. Bayes’ Theorem                          | 8        |
| **Random Variables**                        |          |
| 17. Random Variable                         | 6        |
| 18. Cumulative Distribution Function        | 6        |
| 19. Inverse Cumulative Distribution Function| 5        |
| 20. Probability Density Function            | 3        |
| 21. Probability Function                    | 4        |
| 22. Probability Mass Function               | 2        |
| 23. Quantile Function                       | 5        |
| 24. The Bernoulli Distribution              | 3        |
| 25. The Binomial Distribution               | 9        |
| 26. The Discrete Uniform Distribution       | 3        |
| 27. The Geometric Distribution              | 1        |
| 28. The Point Mass Distribution             | 1        |
| 29. The Poisson Distribution                | 10       |
| 30. t Distribution                          | 1        |
| 31. Cauchy Distribution                     | 1        |
| 32. Exponential Distribution                | 2        |
| 33. Gamma Distribution                      | 5        |
| 34. Gaussian                                | 7        |
| 35. Normal                                  | 7        |
| 36. The $\chi^2$                            | 1        |
| 37. The Beta Distribution                   | 4        |
| 38. The Uniform Distribution                | 1        |
| 39. Bivariate Probability Density Function  | 4        |
| 40. Joint Mass Function                     | 1        |
| 41. Marginal Density Function               | 4        |
| 42. Marginal Mass Function                  | 2        |
| 43. Independent Random Variables            | 5        |
| 44. Conditional Probability Density Function| 5        |
| 45. Conditional Probability Mass Function   | 3        |
| 46. Independent And Identically Distributed | 1        |
| 47. Multinomial                             | 7        |
| 48. Multivariate Normal                     | 1        |
| **Expectation**                             |          |
| 49. Expected Value                          | 22       |
| 50. First Moment                            | 1        |
| 51. Mean                                    | 1        |
| 52. Covariance And Correlation              | 2        |
| 53. Variance                                | 8        |
| 54. Conditional Expectation                 | 9        |
| 55. Conditional Variance                    | 2        |
| 56. Moment Generating Function              | 11       |
| **Inequalities**                            |          |
| 57. Chebyshev’S Inequality                  | 7        |
| 58. Hoeffding’S Inequality                  | 2        |
| 59. Markov’S Inequality                     | 3        |
| 60. Mill’S Inequality                       | 1        |
| 61. Cauchy-Schwartz Inequality              | 1        |
| 62. Jensen’S Inequality                     | 2        |
| **Convergence of Random Variables**         |          |
| 63. Converges To $X$ In Distribution        | 2        |
| 64. Converges To $X$ In Probability         | 2        |
| 65. Converges To $X$ In Quadratic Mean      | 1        |
| 66. The Weak Law Of Large Numbers           | 2        |
| 67. The Central Limit Theorem               | 2        |
| 68. The Delta Method                        | 1        |
| 69. The Multivariate Delta Method           | 2        |

Table 5: 69 Background knowledge concepts. A few concepts are synonymous, for example 34 and 35.
| Concept                          | Train | #. Dev | Test |
|---------------------------------|-------|--------|------|
| The Binomial Distribution       | 78    | 8      | 13   |
| Conditional Probability         | 135   | 16     | 25   |
| Correlation                     | 141   | 15     | 26   |
| Covariance                      | 64    | 8      | 7    |
| Expected Value                  | 134   | 16     | 21   |
| Independent Events              | 73    | 7      | 8    |
| Normal Distribution             | 193   | 22     | 29   |
| Probability Density Function    | 49    | 4      | 3    |
| The Poisson Distribution        | 58    | 6      | 9    |
| Random Variable                 | 133   | 20     | 26   |
| Variance                        | 103   | 14     | 17   |
| **Total**                       | **1,161** | **136** | **184** |

| #. Unique Problems              | 904   | 110    | 157  |
| #. Unique Answers               | 904   | 110    | 157  |
| **Total**                       | **1808** | **220** | **314** |

Table 6: Per concept listing of problems in the train/dev/test partitions. Some problems can be categorized under multiple concepts. Each problem has a corresponding accepted answer.

| TAG                  | Excluded/Ignored | REASON                      |
|----------------------|------------------|------------------------------|
| Matlab               | Excluded         | programming problems        |
| R                    | Excluded         | programming problems        |
| Probability          | Ignore           | Too Generic                 |
| Mathematical-statistics | Ignore        | Too Generic                 |
| Meta-analysis        | Ignore           | Too Generic                 |
| Hypothesis-testing’  | Ignore           | Too Generic                 |
| Distributions        | Ignore           | Too Generic                 |
| Self-study           | Ignore           | Too Generic                 |
| Intuition            | Ignore           | Too Generic                 |
| Definition           | Ignore           | Too Generic                 |

Table 7: stats.stackexchange data dump pre-processing: excluded and ignored tags. If the tag is excluded, we remove from our dataset problems with this tag. If a tag is ignored, we can keep the problem but only if it has other tags besides an ignored tags.
| Method | Dev F1     |
|--------|-----------|
| MaxEnt | 0.528 ± 0.003 |
| MLP    | 0.510 ± 0.020 |
| LSTM   | 0.474 ± 0.027 |
| GRU    | 0.472 ± 0.041 |
| CNN    | **0.706 ± 0.028** |

Table 8: Multi-label concept classification dev performance average over all 11 seeds: [0-5, 10, 100, 1000, 10000, 1000000]
| Method       | #. Parameters | Epochs (bounds) | Epochs (best) | Number of epochs search trials | Choosing hyperparameter values | Seed |
|--------------|---------------|-----------------|---------------|-------------------------------|-------------------------------|------|
| MaxEnt       | 8,459         | [1-300]         | 300           | 31                            | uniform (1, 10, 20, 30, ...)  | 3    |
| MLP          | 924,683       | [1-300]         | 300           | 31                            | uniform: 1, 10, 20, 30, ...  | 1    |
| LSTM         | 4,469,771     | [1-300]         | 300           | 31                            | uniform: 1, 10, 20, 30, ...  | 2    |
| GRU          | 3,583,499     | [1-300]         | 300           | 31                            | uniform: 1, 10, 20, 30, ...  | 2    |
| CNN          | 3,138,829     | [1-300]         | 20            | 31                            | uniform: 1, 10, 20, 30, ...  | 10000|
|              |               |                 |               |                               | combinations of [1,2,3,4,5,6] |      |
|              |               |                 |               |                               | [3,4,5,6]                     |      |
|              |               |                 |               |                               | manual                        |      |
|              |               |                 |               |                               | [50-300]                      |      |
|              |               |                 |               |                               | [192]                         |      |
|              |               |                 |               |                               | manual                        |      |
| Prototypical Networks | 2,214,146 | [10-200]         | 100           | uniform                      |                                |      |
|              |               |                 |               | support size                 | 10                            |      |
|              |               |                 |               | query size                   | 15                            |      |

Table 9: Hyperparameter search for methods in Table 2 of the paper.
7 Experimental Evaluation

Computing Infrastructure. Operating System: 16.04.1-Ubuntu/ x86_64. GPU: GeForce GTX 1080 Ti with 12 GB of memory.

Evaluation Metrics. Evaluation metric is Macro f1 score: We used the implementation in sklearn. In python, the code is: `from sklearn.metrics import f1_score`

Training and Method details The GCN input features are 768-dimensional, and its hidden states are 100 dimensional, and the number of convolutional layers is 3.

We use the ADAM optimizer, and dropout 0.2. Word embeddings are initialized with pretrained 768 dimensional BERT embeddings. The MLP has 3 hidden layers, and the dimensionalities of hidden states is 512. The LSTM is a one layer bidirectional LSTM, with hidden dimensions 384. The GRU is a one layer bidirectional GRU.

Parameter details for the methods for unknown extraction are in Table 10.
| Method            | #. Parameters | Epochs (bounds) | Epochs (best) | Number of epochs search trials | Choosing epochs values | Seed | Filter Sizes (bounds) | Filter Sizes (best) | Method of choosing Filter Sizes | Filters (bounds) | Filters (best) | Method of choosing Filters |
|-------------------|---------------|-----------------|---------------|-------------------------------|------------------------|------|-----------------------|-------------------|---------------------------------|-----------------|---------------|------------------------|
| MaxEnt            | 1,537         | [1-100]         | 30            | 11                            | uniform (1, 10, 20, 30, ...) | 3    |                       |                   |                                 |                 |               |                        |
| MLP               | 1,312,769     | [1-50]          | 30            | 6                             | uniform (1, 10, 20, 30, ...) | 10   |                       |                   |                                 |                 |               |                        |
| CNN               | 1,414,275     | [1-100]         | 20            | 6                             | uniform (1, 10, 20, 30, ...) | 4    | combinations of [1,2,3,4,5,6] | [1,2]             | manual                          | [50-300]        | [192]         | manual                 |
| CNN_NoContext    | 1,166,467     | [1-50]          | 30            | 6                             |                         | 0    |                       |                   |                                 |                 |               |                        |
| CNN+Graph         | 1,502,386     | [1-50]          | 30            | 6                             |                         | 0    |                       |                   |                                 |                 |               |                        |
| CNN+Graph+LSTM    | 2,980,018     | [1-50]          | 30            | 6                             |                         | 10000|                       |                   |                                 |                 |               |                        |
| CNN+Graph+GRU     | 2,610,610     | [1-50]          | 30            | 6                             |                         | 10   |                       |                   |                                 |                 |               |                        |

Table 10: Hyperparameter search for methods in Table 3 of the paper.