Machine comprehension using semantic graphs

Kinga Andrea Gémes   Ádám Kovács   Gábor Recski

Department of Automation and Applied Informatics
Budapest University of Technology and Economics

kinga.andrea.gemes@gmail.com
kovacs.adam@aut.bme.hu
recski@aut.bme.hu

Abstract. We present a novel method for recognizing entailment using semantic graphs and apply it to the `2018 Semeval task on Machine Comprehension (MC)`. Concept graphs are built automatically from MC texts, questions, and answers, using the `4lang` system [1]. A strong baseline method using only these graphs is presented, followed by an enhancement of a state-of-the-art system [2]. Preliminary results suggest that these features achieve a .5 percentage point improvement over the original system [3].

Keywords: Semantic parsing; Natural language processing; 4lang; Graph transformation; Comprehension

1 Introduction

Explicit representations of natural language semantics are rarely used in state of the art systems for popular semantics tasks such as measuring semantic similarity or machine comprehension. Virtually all systems competing at popular challenges (e.g. [4, 5]) rely on word embeddings as the sole representation of word meaning. Methods using graphical representation have been already shown to be capable of improving state of the art system on the task of measuring semantic similarity of pairs of English words [6]. In this paper we use similar graphs as simple but powerful tools for measuring textual entailment, which in turn allows us to improve a top system on the Machine Comprehension task at Semeval 2018.

Section 2.1 provides an overview of the machine comprehension task at SemEval 2018 and the top-ranking system that we later improve using our graph-based method. In Section 2.2 we also briefly describe the semantic parser `4lang` and some of its previous applications. Section 3 presents a simple baseline method for measuring textual entailment and its application to the comprehension task. Section 3.1 reports the results of applying the baseline method to the MC task and also of using it as an extra feature in the neural network based `Yuanfudao` system.

2 Background

2.1 Machine comprehension

The 2018 Semeval Task *Machine comprehension using commonsense knowledge*[1] requires participants to train systems that can choose the correct answer to simple multiple choice questions based on short passages describing simple chains of events. Data for both training and testing is extracted from the `MCScript` dataset [7]. Some questions can only be answered using commonsense knowledge, and are explicitly labeled as such. For example, one passage might describe a story of a gardener planting a tree, and one of the questions would subsequently ask whether the gardener used his hands or a shovel to dig a hole for the tree, even though the answer to this question is not present in the passage. The top two systems, `HFL-RC` [8] and `Yuanfudao` [2] achieved accuracy scores of 84.15% and 83.95% on the test data, respectively. In our experiments we used semantic graphs to augment the `Yuanfudao` system, since its source code is publicly available[2] and since it already employs successfully a knowledge base representing semantic relationships among pairs of words.

The `Yuanfudao` system implements a *Three-way Attentive Network* (TriAN), an ensemble of three LSTMs augmented with various attention mechanisms, to model for each question interactions between question, possible answers, and the passage that may or may not contain the correct answer to the question. An overview of the original system is reproduced in Figure 1.

Input features to the LSTM include word, part-of-speech, and NER embeddings, but also feature vectors representing semantic relationships between pairs of words according to the external knowledge base `ConceptNet` [9]. This database encodes, among many other links, a small set of semantic relations between word pairs such as `IsA`, `UsedFor`, or `CapableOf`. One component of a passage's input representation in the `Yuanfudao` system is a vector that encodes for each word of the passage, via a 10-dimensional embedding, the `ConceptNet` relation between it and some word of an answer candidate (if multiple words in the answer are connected to the given word, one is picked at random). It is this representation that we will replace/augment in our experiments using the simple graph-based entailment metric defined in Section 3.

2.2 4lang

The `4lang` system of semantic representation [10] represents the meaning of linguistic units (both words and phrases) as directed graphs of syntax-independent concepts. Those representing binary relations are connected to their arguments via edges labeled `1` and `2`, all other relations are treated uniformly: `0`-edges represent attribution ($\texttt{dog} \xrightarrow{0} \texttt{large}$), hypernymy ($\texttt{dog} \xrightarrow{0} \texttt{mammal}$) and unary predication ($\texttt{dog} \xrightarrow{0} \texttt{bark}$). Concepts have no grammatical attributes and no

[1] `https://competitions.codalab.org/competitions/17184`

[2] `https://github.com/intfloat/commonsense-rc`

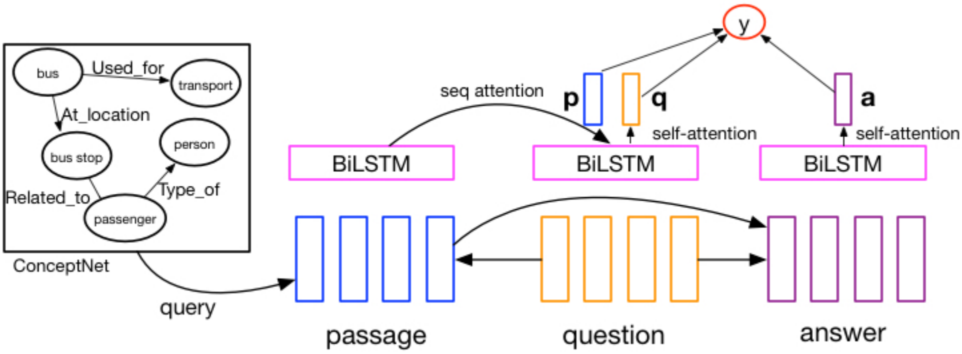

Figure 1: Structure of the original network [2, p.2]

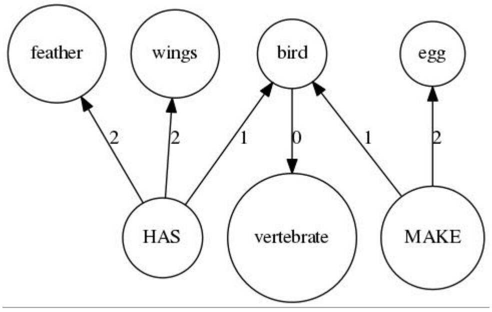

Figure 2: 4lang definition of `bird`.

event structure, e.g. the phrases *water freezes* and *frozen water* would both be represented as *water* $\xrightarrow{0}$ *freeze*.

The example in Figure 2 shows the `4lang` definition of the concept `bird`. This definition was built manually, as part of the `4lang` dictionary [11], but similar definitions have been created automatically from definitions of monolingual dictionaries such as Longman, using the `dict_to_4lang` tool [1].

The open-source 4lang pipeline[3] contains tools for generating directed graphs from raw text by mapping dependency edges in the output of the Stanford parser [12] to `4lang` subgraphs over concepts corresponding to each word of the original sentence. Optionally, the `4lang` system allows us to *expand* graphs, a process which unifies the graph with the definition graphs of each concept. Besides being an open-source software library, the `4lang` parser is also accessible via a public REST API at `http://hlt.bme.hu/4lang`.

Graphs generated by the `4lang` parser have previously been used successfully in measuring semantic similarity. The current state of the art system on the `SimLex-999` benchmark [13] outperforms previous top systems by utilizing a simple similarity metric between `4lang` definitions of pairs of English words [6]

[3] `https://github.com/kornai/4lang`

2.3 Comprehension, entailment, and knowledge bases

In the next section we shall present a simple method for measuring *support*, the continuous counterpart of *entailment*, based on graphical representations of meaning, and use this metric in a baseline for machine comprehension and to improve a state of the art MC system. Although explicit representations of semantics are rarely used for this purpose, in recent years there have been several attempts at leveraging lexical ontologies in machine comprehension, and the approach of using textual entailment as an intermediary task is also not new. [14] achieves competitive results on the `MCTest` dataset [15] by generating answer candidates and ranking them using a separate RTE system, which is trained on the Stanford Natural Language Inference (SNLI) dataset [16] but also relies on an explicit measure of lexical overlap between sentence pairs. Other recent systems are various extensions of a baseline proposed by [15] that measures a weighted overlap between pairs of bag-of-words representations, e.g. [17] applies the frame semantic parser of [18] and includes features representing overlap between bag-of-frame and bag-of-argument representations. Finally, the `Yuanfudao` system presented in this section is the most recent example of enhancing the performance of an MC system using a lexical knowledge base: ablation studies show that their top-ranking accuracy score of 83.84% drops to 82.78% if `ConceptNet`-based features are removed.

3 Method

We define a simple metric between pairs of `4lang` graphs that we intend to use for measuring entailment between a paragraph and a sentence. We shall define the degree to which some graph G_1 *supports* another graph G_2 as the ratio of edges in G_2 that are also present in G_1:

$$S(G_1, G_2) = \frac{|E(G_1) \cap E(G_2)|}{|E(G_2)|}$$

Two (directed) edges are identical if their source and target nodes and their label are all identical. Based on early findings we used only *expanded* `4lang` graphs (see Section 2.2) for measuring support. To create a simple baseline solution for the Machine Comprehension task, we compare answer candidates to each question by comparing the degree of support for each in the passage, based on the `4lang` representations of each piece of text. For wh-questions we can create representations of each answer by merging the question graph's wh-node with the graph of each answer graph (see Figure 3). Our baseline method will simply pick the answer candidate with the higher support score.

3.1 Experiments

Demonstrating the algorithm behind our baseline, let's look at the example passage:

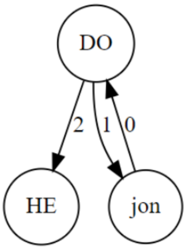

Figure 3: Merged graph of answer candidate *"John"* for the question *Who did it?*

> *"I went into my bedroom and flipped the light switch. Oh, I see that the ceiling lamp is not turning on. It must be that the light bulb needs replacement. I go through my closet and find a new light bulb that will fit this lamp and place it in my pocket. I also get my stepladder and place it under the lamp. I make sure the light switch is in the off position. I climb up the ladder and unscrew the old light bulb. I place the old bulb in my pocket and take out the new one. I then screw in the new bulb. I climb down the stepladder and place it back into the closet. I then throw out the old bulb into the recycling bin. I go back to my bedroom and turn on the light switch. I am happy to see that there is again light in my room."*

And a question related to the text: *Which room did the light go out in?* and the answers:

- *"Kitchen."*
- *"Bedroom."*

First we build the expanded graph from the text. After we build the merged graphs (for the demonstration, we now only build the graphs without expansion) seen in Figure 4. After the merging, we compare both of the graphs to the passage graph applying our defined metric.

We tested the baseline method described in the previous section on a subset of all questions in the train section of the MC dataset: wh-questions that were not categorized as “common-sense”. Of this subset of 5,375 questions (of a total of 9,731), our method correctly answers 3,671, achieving an accuracy score of 68.3%. We then proceeded to use the metric underlying our baseline as an additional feature in the `Yuanfudao` system.

The most straightforward way of incorporating our metric into the system introduced in Section 2.1 is by creating vectors similar to those representing ConceptNet relations between words of a passage and words in each answer candidate. Since these vectors represent word-to-word relationships, we measure the support between pairs of `4lang` definition graphs, and for each word in the passage we take the maximum support score over all words of the answer

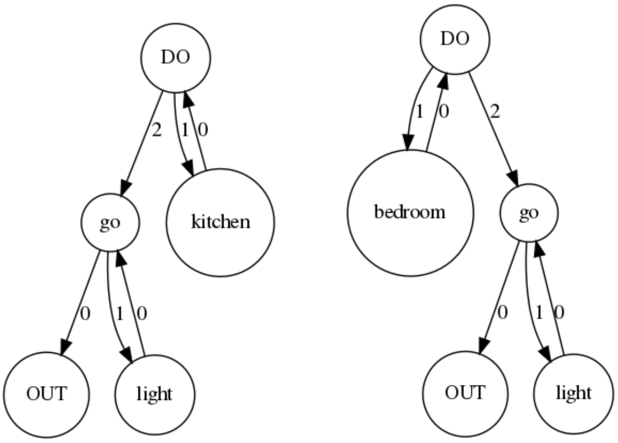

Figure 4: Merged graphs for answer candidates *"Kitchen."* and *"Bedroom."*

candidate. Elements of a vector for a passage P and a possible answer A are hence defined as:

$$S_i^{(P,A)} = \max_{A_j \in A} S(P_i, A_j)$$

We compare the original TriAN system with two setups: in one case, the input representation of each passage is extended with the 4lang-based vector defined above. In the second setup, the 4lang-based vector is added and the ConceptNet-based relation embedding is removed. Hyperparameters of training are taken from the original system, they have not been tuned further. The original system benefitted from pretraining the model on the RACE dataset [19]. Table 1 summarizes our results without this pretraining step and Table 2 contains the accuracy we achieved on the pretrained models.

The 4lang-based support scores achieve a ca. 0.5 percentage point improvement over the original TriAN configuration, consistent across development and test datasets. Effects of each component on accuracy, as measured here without pretraining the model, are not in line with the findings of the ablation study of [2], as they suggest that ConceptNet features can be discarded without a drop in performance. Further experiments are required to explore how our enhancements can improve further the top-ranking system that employs pretraining and an ensemble of multiple models. Our results so far on the pretrained, 9-way ensembled models are summarized in Table 3.

We also plan to incorporate sentence-level support into the system as a more direct application of our baseline.

model	dev	test
TriAN, no ConceptNet	82.8%	80.2%
TriAN, with ConceptNet	82.7%	80.5%
TriAN, with 4lang	**83.2%**	**80.9%**
TriAN, with both	83.1%	80.8%

Table 1: Effect of `4lang` and `ConceptNet` on results

model	dev	test
TriAN, no ConceptNet	83.7%	81.9%
TriAN, with ConceptNet	82.5%	80.3%
TriAN, with 4lang	84.2%	81.5%
TriAN, with both	**83.4%**	**82.9%**

Table 2: Effect of `4lang` and `ConceptNet` on the pretrained models

pretrained, ensembled model	test
TriAN, no ConceptNet	82.95%
TriAN, with ConceptNet	83.697%
TriAN, with 4lang	82.8%
TriAN, with both	**83.73%**

Table 3: The effect of `4lang` and `ConceptNet` on the pretrained and ensembled models

Acknowledgments

Research partially supported by National Research, Development and Innovation Office NKFIH grant #120145 and by National Excellence Programme 2018-1.2.1-NKP-00008: Exploring the Mathematical Foundations of Artificial Intelligence.

References

[1] G. Recski, "Building concept graphs from monolingual dictionary entries," in *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)* (N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, eds.), (Portorož, Slovenia), European Language Resources Association (ELRA), 2016.

[2] L. Wang, M. Sun, W. Zhao, K. Shen, and J. Liu, "Yuanfudao at semeval-2018 task 11: Three-way attention and relational knowledge for commonsense machine comprehension," *arXiv preprint arXiv:1803.00191*, 2018.

[3] K. Gémes and A. Kovács, "Semantic parsing with graph transformations," tech. rep., Budapest University of Technology and Economics, Budapest, 2018.

[4] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, (Vancouver, Canada), pp. 1–14, Association for Computational Linguistics, 2017.

[5] J. Camacho-Collados, M. T. Pilehvar, N. Collier, and R. Navigli, "Semeval-2017 task 2: Multilingual and cross-lingual semantic word similarity," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, (Vancouver, Canada), pp. 15–26, Association for Computational Linguistics, 2017.

[6] G. Recski, E. Iklódi, K. Pajkossy, and A. Kornai, "Measuring semantic similarity of words using concept networks," in *Proceedings of the 1st Workshop on Representation Learning for NLP*, (Berlin, Germany), pp. 193–200, Association for Computational Linguistics, 2016.

[7] S. Ostermann, A. Modi, M. Roth, S. Thater, and M. Pinkal, "MCScript: A Novel Dataset for Assessing Machine Comprehension Using Script Knowledge," in *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, (Miyazaki, Japan), 2018.

[8] Z. Chen, Y. Cui, W. Ma, S. Wang, T. Liu, and G. Hu, "Hfl-rc system at semeval-2018 task 11: Hybrid multi-aspects model for commonsense reading comprehension," *arXiv preprint arXiv:1803.05655*, 2018.

[9] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge.," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pp. 4444–4451, 2017.

[10] A. Kornai, J. Ács, M. Makrai, D. M. Nemeskey, K. Pajkossy, and G. Recski, "Competence in lexical semantics," in *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, (Denver, Colorado), pp. 165–175, Association for Computational Linguistics, 2015.

[11] A. Kornai and M. Makrai, "A 4lang fogalmi szótár," in *IX. Magyar Számitógépes Nyelvészeti Konferencia* (A. Tanács and V. Vincze, eds.), pp. 62–70, 2013.

[12] M.-C. DeMarneffe, W. MacCartney, and C. Manning, "Generating typed dependency parses from phrase structure parses," in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, vol. 6, (Genoa, Italy), pp. 449–454, 2006.

[13] F. Hill, R. Reichart, and A. Korhonen, “Simlex-999: Evaluating semantic models with (genuine) similarity estimation,” *Computational Linguistics*, vol. 41, no. 4, pp. 665–695, 2014.

[14] B. Wang, S. Guo, K. Liu, S. He, and J. Zhao, “Employing external rich knowledge for machine comprehension.,” in *IJCAI*, pp. 2929–2925, 2016.

[15] M. Richardson, C. J. Burges, and E. Renshaw, “Mctest: A challenge dataset for the open-domain machine comprehension of text,” in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 193–203, 2013.

[16] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, “A large annotated corpus for learning natural language inference,” in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, (Lisbon, Portugal), pp. 632–642, Association for Computational Linguistics, 2015.

[17] H. Wang, M. Bansal, K. Gimpel, and D. McAllester, “Machine comprehension with syntax, frames, and semantics,” in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, vol. 2, pp. 700–706, 2015.

[18] D. Das, N. Schneider, D. Chen, and N. A. Smith, “Probabilistic frame-semantic parsing,” in *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, pp. 948–956, Association for Computational Linguistics, 2010.

[19] G. Lai, Q. Xie, H. Liu, Y. Yang, and E. Hovy, “RACE: Large-scale reading comprehension dataset from examinations,” *arXiv preprint arXiv:1704.04683*, 2017.