Machine comprehension using semantic graphs

Anonymous CoNLL submission

Abstract

This paper presents our system for Semeval-2018, where we represent each text, question and answers as directed concept graphs. We merge each question-answer pair graphs, and calculate assymetric jaccard similarities between them. Our system achieves 68,3% accuracy score, which reflects a strong baseline.

We present a set of pilot experiments for augmenting a generic, open-domain knowledge base using a graph-based lexical ontology of English and simple inference rules. The WikiData knowledge-base contains facts encoded as relation triplets, such as author(George Orwell, 1984), based on which naive speakers can easily establish additional facts such as that George Orwell is a person and 1984 is some written work, most likely a book. To automate this type of inference we need models of lexical semantics that are more explicit than the distributional models commonly used in computational semantics. The 4lang library provides tools for building concept graph representations of the semantics of natural language text, its module `dict_to_4lang` processes entries of monolingual dictionaries to build 4lang-style definition graphs of virtually any English word. The representation of "author" will likely contain edges corresponding to facts such as `IS_A(author, person)` and `write(author, book)`. We define simple templates that use these representations for inference over WiktData facts; our method yields millions of new facts with high accuracy (over 90% according to manual evaluation)

1 Introduction

The main goal of a Knowledge base population system is to discover facts about entities and augment a knowledge base with these facts, this is done through multiple tasks. Generally these tasks are hybrid systems where they combine various aspects of Question Answering (passage retrieval) and Information Extraction (answer extraction) (**?**). However these systems diverge from our system because their motivation is to automatically obtain information from news and unstructured data (**?**), while we work with structured knowledge base and we apply inference rules for augmentation.

In this paper we present a strong baseline for SemEval-2018 Task Machine comprehension using directed semantic graphs discussed in (Kornai and Makrai, 2013), The main idea behind our solution was to represent the answers and the texts as graphs and calculate graph similarities beetween them.

The method presented yields millions of new triplets, the quality of which we evaluate by manual inspection of ca. 200 of the most common relations. Our system is available on GitHub[1] under an MIT license.

`WikiData`[2] is a public domain knowledge base containing attribute-value type information about more than 30 million entities. For each entity, WikiData contains pairs of *properties* (attribute) and *values*, which may contain pointers to other entities. In case of the entity `1984`, the value of the property `author` is `George_Orwell`. An alternative representation of the dataset is in the form of relational triplets, this would represent the above fact as a single binary relation `author(George_Orwell, 1984)`. We use This latter representation when processing WikiData.

The `4lang` system of semantic representation (Kornai et al., 2015) represents the meaning of linguistic units (both words and phrases) as directed graphs of grammar- and language-independent concepts. Concepts representing binary relations

[1] https://github.com/adaamko/4lang
[2] https://www.wikidata.org/

are connected to their arguments via edges labelled 1 and 2, all other relations are treated uniformly: 0-edges represent attribution (dog $\xrightarrow{0}$ large), hypernymy (dog $\xrightarrow{0}$ mammal) and unary predication (dog $\xrightarrow{0}$ bark). The example in Figure 1 shows the 4lang definition of the concept bird. This definition was built manually, as part of the 4lang dictionary (Kornai and Makrai, 2013), but similar definitions have been created automatically from definitions of monolingual dictionaries such as Longman, using the dict_to_4lang tool (Recski, 2018). We process this set of definition graphs when performing inference over WikiData triplets.

2 Background

2.1 4lang

Under the name *4lang* we mean both a formalism that represents meaning by building directed graphs and a name of a manually built lexicon. [3]. The *4lang* graphs are directed graphs which nodes are the concepts, and the edges have three types. The first type is the most common, the 0-edge, that can stand for attribution (*song* $\xrightarrow{0}$ *music*), the IS_A relation (*father* $\xrightarrow{0}$ *male*) and unary predication (*dog* $\xrightarrow{0}$ *run*). The *4lang* has two more edges defined, namely 1 and 2 type. They connect binary predicates to their arguments (*Peter* $\xleftarrow{1}$ *play* $\xrightarrow{2}$ *football*).

The nodes have no grammatical attributes, so part-of-speech and voice doesn't play part. This means that both *water freezes* and *frozen water* would both be represented with the following edge *water* $\xrightarrow{0}$ *freeze*.

The example in Figure 1 shows the 4lang definition of the concept bird. This definition was built manually, as part of the 4lang dictionary (Kornai and Makrai, 2013), but similar definitions have been created automatically from definitions of monolingual dictionaries such as Longman, using the dict_to_4lang tool (Recski, 2018).

The full 4lang pipeline is available in library[4] that contains tools for generating directed graphs from raw text *text_to_4lang* and *dict_to_4lang* that parses dictionary entries. The main module of the library is the *dep_to_4lang* that maps the output of the Stanford parser to the directed graphs (DeMarneffe et al., 2006).

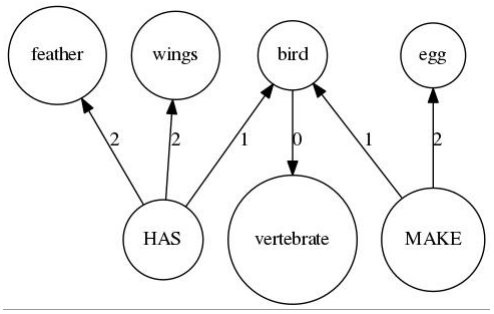

Figure 1: 4lang definition of bird.

3 Machine comprehension

The Machine Reading Comprehension has been a very hot topic in recent natural language processing field. The task contains narrative texts that happens in everyday activities and the system requires to answer multiple-choice questions with the information described in the texts. The questions are associated with two short answers with limited length, described in (cod, 2018). The texts cover various number of everyday scenarios.

The state-of-the-art system (Zhipeng Chen, 2018) uses Hybrid Multi-Aspects model, that mimic the human's intuitions on dealing with multiple-choice. They calculate attention among text and question and answers. They use an Embedding layer that project text, question and answers into embedding representations with three components: Word embedding, Char embedding and Feature embedding. They describe a RNN Layer with Bi-directional LSTM. After that they defined an Attention layer to calculate attentions between different combinations. They achieved state-of-the-art result with accuracy score of 84.13%. In paper described in (Liang Wang, 2018) uses Three-way Attentive Networks to model the interactions between the text and question-answers. They chose attention mechanism for reading comprehension. Their system uses word-level attention and only one layer of LSTM. The system achieved accuracy score of 83.95%.

3.1 4lang baseline

The 4lang system can be directly accessed as a Restful service available [5]. Our system aims to

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

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

[5] http://4lang.hlt.bme.hu

model the MRC task with directed graphs. First we need to define how can we calculate similarities between graphs. For this task we introduce assymetric similarity between two concepts. This metric is based on a intuition that similar concepts will contain similar edges in their definition graphs. Let us look at the two sentences

- My poor wife! (P)
- I feel bad for my wife. (H)

We can have an assumption that the second sentence follows from the first sentence. If we want to model this problem, we can generate the concept graphs with the sentences. The generated graph from the premise sentence can bee seen in Figure 2, and the hypothesis in Figure 3. If the graphs meaning corresponds to the sentences, we can calculate similarites between them as:

$$sim = \frac{|E(P) \cap E(H)|}{|E(H)|}$$

We can achieve higher success rate if we expand our graph with the words definitions. We can build each word's definition graph, and merge the graphs with the original sentence's graph. Staying with our examples, lets say we have the word *wife*'s definition graphs with the following edge *wife* $\xrightarrow{0}$ *woman* then we can expand our graph with the given information by essentially merging the two graphs together, this example can be seen in Figure 4.

With the described logic, let's say we have the following question: *Who did it?*, and we have an answer *Jon*, then the question's definition graph will most likely contain an edge corresponding the question word *Who?*. We can merge the question with the answer's definition graph by replacing the question word with the answer's graph (for this first we need to find the ROOT node in the answer graph) as showed in Figure 5. We do this for each question-answer pair, and after that we can calculate the similarities to the text's definition graph, and if we are ready to make an assumption that the higher similarities corresponds if the answer will be True or False, than the more similar question-answer pair will be the True answer, and the other will be the False.

4 Results and evaluation

With the described mechanism now we apply our algorithm to the task. First we need to filter the dataset to contain a question word. This way 5375

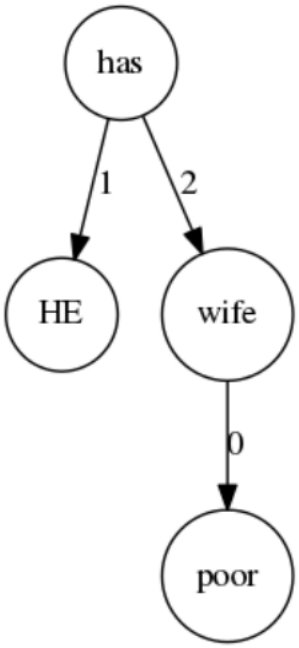

Figure 2: 4lang definition of `My poor wife!`.

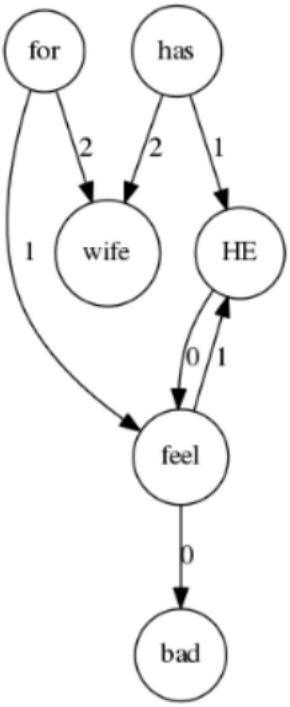

Figure 3: 4lang definition of `I feel bad for my wife`.

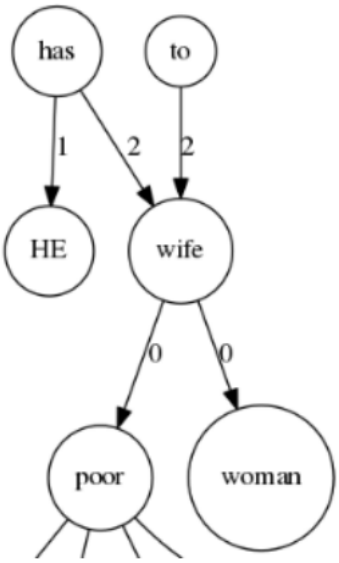

Figure 4: 4lang expanded graph

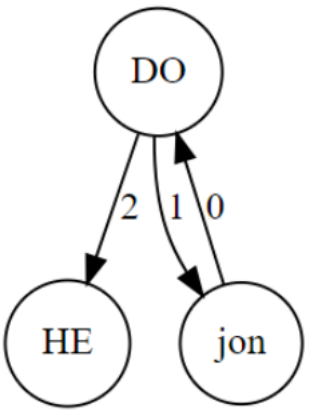

Figure 5: 4lang merged graph

question remaind. For these question our system achieved accuracy of 68.3%.

5 Method

We implement two simple patterns for performing inference using `WikiData` triplets and `4lang` definitions. Given the triplet `author(George_Orwell, 1984)` and the definition graph of `author` in Figure **??**, we should be able to infer all edges in the graph in Figure **??**. This requires us to implement two patterns. Given a triplet $R(X, Y)$, we first find nodes in the `4lang` definition of R that are connected to R by an outgoing 0-edge (e.g. `author` $\xrightarrow{0}$ `write`) and assume that each of these 0-relations holds for X. The second inference we would like to make is `1984` $\xrightarrow{0}$ book based on the `4lang` edge `write` $\xrightarrow{2}$ book. To this end we implement the rule that if for any relation R and concepts B and C we find $R \xrightarrow{0} B \xrightarrow{2} C$, then for each triplet $R(X, Y)$ we add the edge $Y \xrightarrow{0} C$.

An issue we encountered early concerns words with multiple outgoing 0-edges in their definition graph. Often, this is the result of a dictionary definition that lists several categories that the concept may belong to, e.g. the definition of `employer` is *a person, company, or organization that employs people*. In case of a triplet such as `employer(CIA, Mike_Pompeo)`, we would incorrectly infer `CIA` $\xrightarrow{0}$ `person`. Special treatment for such constructions by 4lang and/or our system might handle these cases and make the inference that the CIA is either a person, a company, or an organization, but for the purpose of the present experiment we decided to discard all `WikiData` relations whose `4lang` definition contains more than one outgoing 0-edge.

Other issues are caused by meaningless or erroneous 0-connections in 4lang graphs that are ultimately limitations of the method used by the `dict_to_4lang` system to build these graphs from natural language definitions. The process involves parsing the definitions with a state-of-the-art dependency parser and mapping grammatical relations between pairs of words to configurations of `4lang` edges. In case of a definition such as **flag:** *piece of cloth with a coloured pattern or picture on it that represents a country*, the definition graph will contain the edge `flag` $\xrightarrow{0}$ `piece`. This information is obviously not informative (to say

the least), we consider it an error when evaluating our system. To make a correct inference about flags similar to those in our previous example, a system would need to learn something along the lines of "*piece of X* $\overset{0}{\rightarrow}$ `X`", which is beyond the scope of the current paper. A final common source of false facts concerns words that are used in WikiData in a very different sense than the one defined by the Longman dictionary, the source of `4lang` definitions. One example is the outdated definition of `developer`: *a person or company that makes money by buying land and then building houses, factories etc on it*, which causes our method to erroneously infer that developers are companies.

6 Evaluation

In the WikiData dataset we count 86.3 million triplets using 893 unique predicates. We started our experiment by preprocessing WikiData to discard fragmentary data (triplets with empty positions) and multi-word predicates that do not lend themselves to the simple methods described in the previous section. After these steps our dataset consisted of 195 predicates and 19.6 million triplets, out of which our first inference pattern was applicable to 108 predicates (covering 9.2 million triplets), the second to 27 predicates (covering 1.4 million triplets). After an initial examination of our output we decided to discard further subsets of predicates: we applied our patterns to predicated whose definition graphs had exactly one outgoing 0- or 2-edge and no incoming edges. We shall see that this step results in a considerable increase in overall accuracy. After these steps we proceeded to apply our two patterns: the first one was now applicable to 84 predicates (8.2 million triplets), the second to 25 predicates (0.8 million triplets). This relatively small number of unique predicates allowed us to inspect all of them manually and estimate the quality of all newly extracted facts: if we find that for some predicate, e.g. `father`, we have made inferences based on the template "$X \overset{0}{\rightarrow}$ `father` implies $X \overset{0}{\rightarrow}$ `male`", we assume that each fact inferred using this template is correct, while for erroneous templates we assume that each extracted fact is false. Figures are shown in Table 1. Note that our evaluation was strict in the sense that we judged incorrect all non-informative edges, e.g. $X \overset{0}{\rightarrow}$ `something`.

The pilot system presented in this paper used simple pattern-based methods for combining facts from a knowledge base with linguistic knowledge represented in a lexical ontology. We believe the significance of this experiment lies not in its yield of millions of high-quality facts with which a knowledge base might be extended, but in its demonstration that inference based on linguistic knowledge is a powerful method for enriching any natural language data.

	1-pattern	2-pattern	total
predicates	84	25	109
correct	55	17	72
new facts	8.2 million	0.83 million	9 million
correct	7.6 million	0.74 million	8.3 million
accuracy	0.92	0.89	0.92

Table 1: Evaluation results

References

2018. *Codalab*.

Marie-Catherine DeMarneffe, William MacCartney, and Christopher Manning. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*, volume 6, pages 449–454, Genoa, Italy.

András Kornai, Judit Ács, Márton Makrai, Dávid Márk Nemeskey, Katalin Pajkossy, and Gábor Recski. 2015. Competence in lexical semantics. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics (*SEM 2015)*, pages 165–175, Denver, Colorado. Association for Computational Linguistics.

András Kornai and Márton Makrai. 2013. A 4lang fogalmi szótár. In *IX. Magyar Számitógépes Nyelvészeti Konferencia*, pages 62–70.

Meng Sun Liang Wang. 2018. Three-way attention and relational knowledge for commonsense machine comprehension. *Semeval-2018*.

Gábor Recski. 2018. Building concept definitions from explanatory dictionaries. *International Journal of Lexicography*.

Yiming Cui Zhipeng Chen. 2018. Hybrid multi-aspects model for commonsense reading comprehension. *Semeval-2018*.