Knowledge base population using natural language inference

Ádám Kovács Gábor Recski
Department of Automation and Applied Informatics
Budapest University of Technology and Economics
adaam.ko@gmail.com
recski@aut.bme.hu

Abstract. 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)

Keywords: semantics; inference; natural language processing;

1 Introduction

We present simple methods for combining facts encoded in the generic knowledge base `WikiData` with definitions in the lexical ontology `4lang` to extract new facts using simple patterns of inference. When provided with a piece of information such as that encoded by the WikiData triplet `author(George Orwell, 1984)`, human speakers can easily establish additional facts such as that *George Orwell* is a person and *1984* is some written work, most likely a book. The automation of this type of inference for all occurrences of a single relation such as *author* could be achieved by manually encoding the piece of knowledge that authors are (typically) humans. Since such knowledge is part of the definition of *author*, we can extract them from a lexicon of definition graphs built using the `4lang` library from definitions of monolingual dictionaries. The method presented yields millions of new triplets, the quality of which we evaluate by

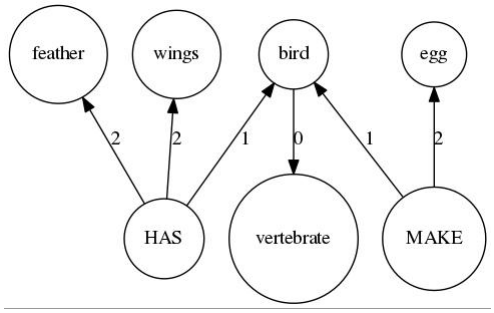

Figure 1: 4lang definition of `bird`.

manual inspection of ca. 200 of the most common relations. Our system is available on GitHub[1] under an MIT license.

2 Background

`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 [1] represents the meaning of linguistic units (both words and phrases) as directed graphs of grammar- and language-independent concepts. Concepts representing binary relations 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 [2], but similar definitions have been created automatically from definitions of monolingual dictionaries such as Longman, using the `dict_to_4lang` tool [3]. We process this set of definition graphs when performing inference over WikiData triplets.

3 Method

We implement two simple patterns for performing inference using `WikiData` triplets and `4lang` definitions. Given the triplet `author(George_Orwell, 1984)`

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

and the definition graph of `author` in Figure 2, we should be able to infer all edges in the graph in Figure 3. 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* $\xrightarrow{0}$ `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.

4 Evaluation

The complete WikiData contains 86.3 million triplets using 893 unique preicates. 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

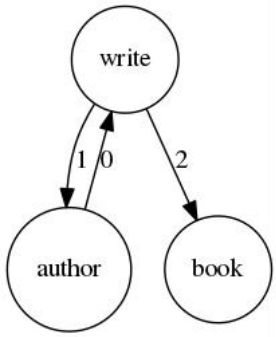

Figure 2: 4lang definition of `author`.

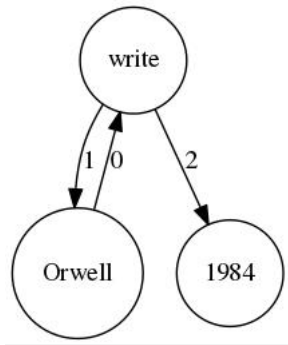

Figure 3: 4lang graph produced from WikiData triplet `author(George_Orwell, 1984)`.

	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

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 pf 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 "$X \xrightarrow{0}$ `father` implies $X \xrightarrow{0}$ `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. $X \xrightarrow{0}$ `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.

References

[1] 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.

[2] 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.

[3] G. Recski, "Building concept definitions from explanatory dictionaries," *International Journal of Lexicography*, to appear.