MyTitle

Adam Kovacs
MyAdress

Abstract

MyAbstract
Keywords: semantics, lexicon, knowledge representation

1. Introduction

In this paper I present a way of matching `WikiData` relations with arguments of `4lang` definitions. The `dict_to_4lang` tool automatically builds graphs from longman dictionary definitions. The full pipeline is available for download under an MIT license at `http://github.com/kornai/4lang`. `WikiData` is a publicly available knowledge base and we can make triplets out of it in the form of predicate(argument1, argument2). If we can make an assumption that these arguments corresponds to each other and a set of patterns can be applied to them, then we can convert a large amount of information from WikiData to the 4lang format and combine the two knowledge.

2. Combine WikiData and 4lang

The 4lang pipeline maps the output of the Standford dependency parser to subgraphs representing the words of each definition. For example `father` is defined in longman as `male parent`. The dictto4lang tool uses this definition to build a 4lang graph seen in Figure 1. If we have a triplet coming from the WikiData knowledge base such as father(Az-Zahir Ghazi, Saladin) and we are ready to make an assumption that the second argument corresponds with the only IS A relation of out graph, then we can combine the fact with the longman definiton to obtain a new graph shown in Figure 2. We have a new machine a IS A relation, the `Saladin` $\xrightarrow{0}$ `male` 0 edge that wasnt present before. We can see that we could obtain a completely new information which was unknown from the definition graph and from the Wikidata alone, and could only be present from the combination of the two. If we want to build 4lang graphs automatically from WikiData, we will require a method for matching these relations, as in the case above. The result will have to be reviewed, and only the reasonable ones have to be selected. If we can apply patterns to these triplets and definitons, we can have a large amount of information retrieved from the combination of the two.

3. Methods

From the examination of the WikiData triplets, we can have a suspicion that if we say have a 0 edge in our definition graph `predicate` $\xrightarrow{0}$ `X` then in our new graph coming from the combination of the WikiData and the 4lang graph, a machine looks like `arg2` $\xrightarrow{0}$ `X` most likely going to have a place. And if we have an edge `predicate` $\xrightarrow{2}$ `X` in our original graph, then we will have an edge `arg1` $\xrightarrow{0}$ `X` in the newly constructed graph. As

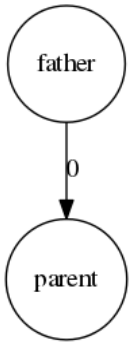

Figure 1: 4lang definition of `father`.

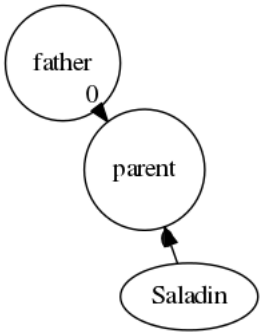

Figure 2: 4lang definition of `fathernew`.

we can see in Figure 3 and 4 new machines appeared such as `OpenCart` $\xrightarrow{0}$ `thing` and `Daniel Kerr` $\xrightarrow{0}$ `made` both of these appear to be valid information thanks to out pattern. Of course this is an ideal situation, this will not be always the case, there are many factors to be considered, when we apply these patterns to out data. We have to take into account the fact, that the triplets coming from the WikiData are not always going to be valid information. This case can be seen in Figure 5 and Figure **??**, where one of the arguments of a WikiData triplet was `novalue`, so the edge created from the triplet does not holds any information. There are cases, when the originally created graph is not completely parsed right from the definition. `flag` is definition in longman is: `piece of cloth with a coloured pattern or picture on it that represents a country`. The definition graph built from this definition is in Figure 7. The machine `flag` $\xrightarrow{0}$ `piece` obviously does not contain valid information, so the triplet flag(Belgium,flag of Belgium) with our current pattern would not add additional information to it. Our parser does not handle when there are multiple choices

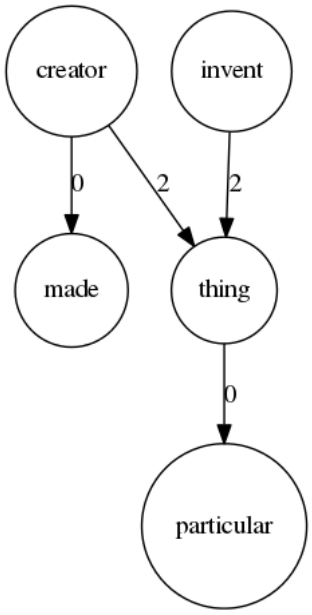

Figure 3: 4lang definition of `creator`.

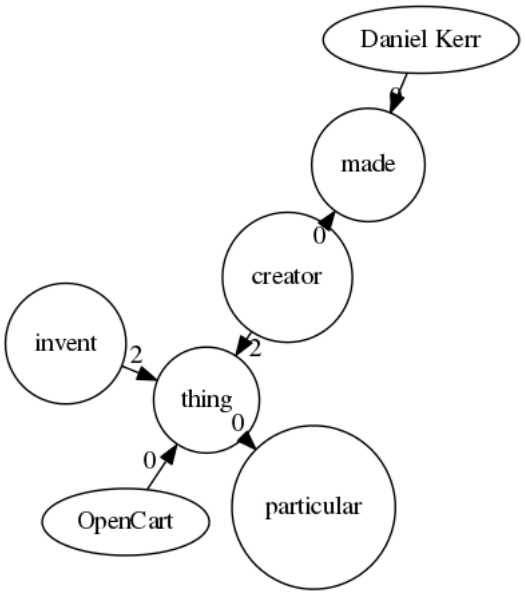

Figure 4: 4lang definition of `creatornew`.

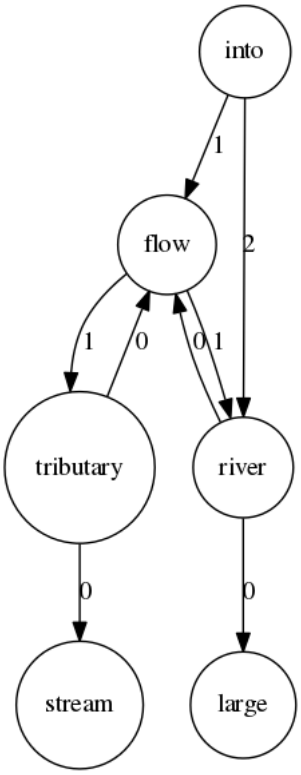

Figure 5: 4lang definition of `tributary`.

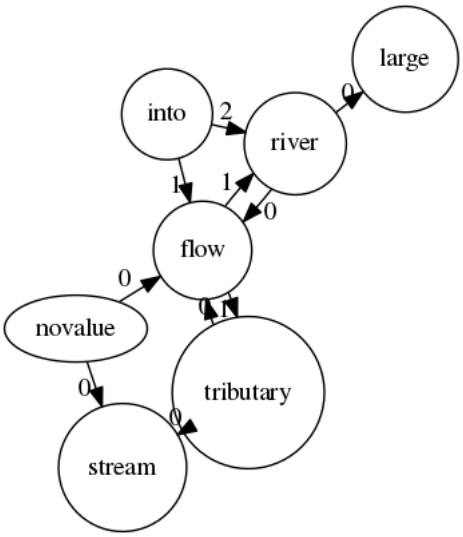

Figure 6: 4lang definition of `tributarynew`.

in a definition. For example longman defines employer `a person, company, or organization that employs people`. The graph constructed in Figure 8 and 9. We have an IS a edge `Central Intelligence Agency` $\xrightarrow{0}$ `person` which we can presume is not a valid assumption, it would be rather a company. The next case, where out pattern can fail is when the WikiData and the longman has different definition of a word, it was the case when we examined the word Developer, which definition in longman was `a person or company that makes money by buying land and then building houses, factories etc on it` but the triplet in WikiData assumed it was a Software Developer as we can see Developer(De Blob, Blue Tongue Entertainment). s

4. The `4lang` formalism

5. Building definition graphs

6. Issues

7. Evaluation

8. Expansion

9. Applications

10. Acknowledgements

The author wishes to thank András Kornai, Márton Makrai, Dávid Nemeskey, and two anonymous reviewers for their

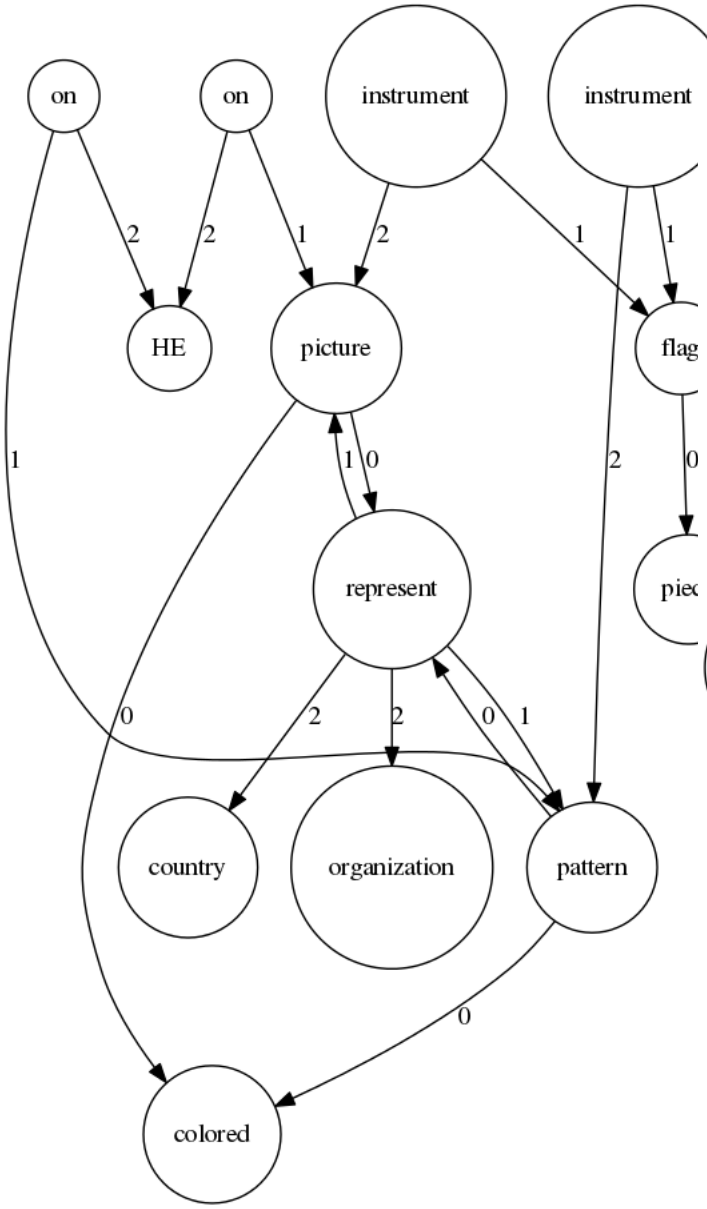

Figure 7: 4lang definition of `flag`.

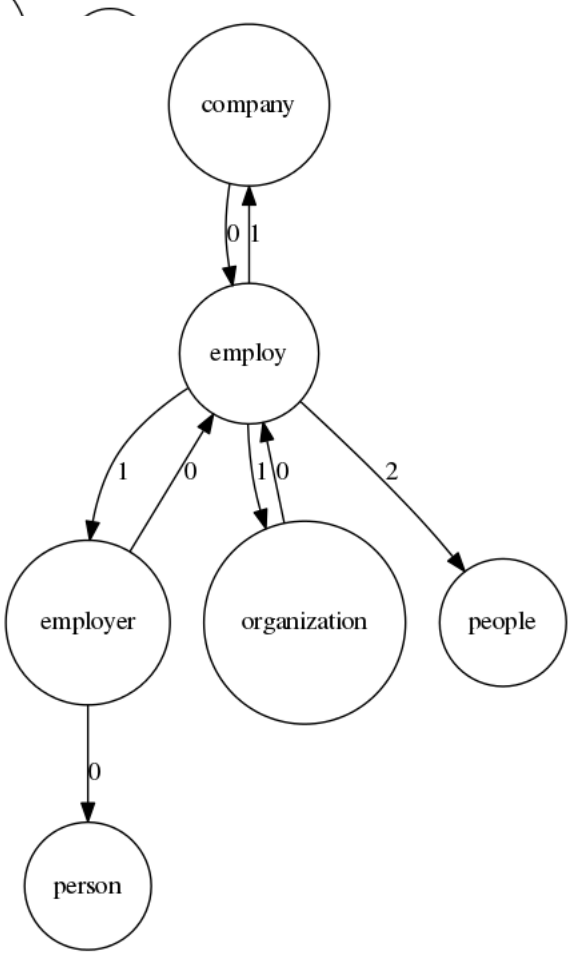

Figure 8: 4lang definition of `employer`.

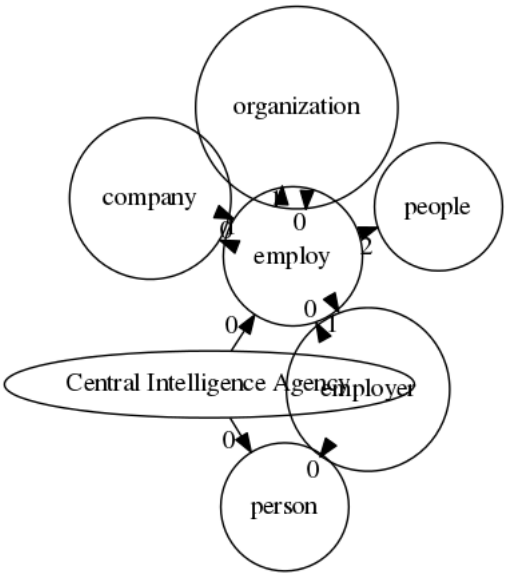

Figure 9: 4lang definition of `employernew`.

many useful comments on earlier versions of this paper.

11. Bibliographical References