Building Concept Graphs from Monolingual Dictionary Entries

Gábor Recski
Research Institute for Linguistics
Hungarian Academy of Sciences
Benczúr u. 33
1068 Budapest, Hungary
recski@mokk.bme.hu

Abstract

We present the dict_to_4lang tool for processing entries of three monolingual dictionaries of English and mapping definitions to concept graphs following the 4lang principles of semantic representation introduced by (Kornai, 2010). 4lang representations are domain- and language-independent, and make use of only a very limited set of primitives to encode the meaning of all utterances. Our pipeline relies on the Stanford Dependency Parser for syntactic analysis, the dep_to_4lang module then builds directed graphs of concepts based on dependency relations between words in each definition. Several issues are handled by construction-specific rules that are applied to the output of dep_to_4lang. Manual evaluation suggests that ca. 75% of graphs built from the Longman Dictionary are either entirely correct or contain only minor errors. dict_to_4lang is available under an MIT license as part of the 4lang library and has been used successfully in measuring Semantic Textual Similarity (Recski and Ács, 2015). An interactive demo of core 4lang functionalities is available at http://4lang.hlt.bme.hu.

Keywords: semantics, lexicon, knowledge representation

1. Introduction

We present the dict_to_4lang tool for automatically building graphs in the style of the 4lang concept dictionary (Kornai et al., 2015) using entries from various explanatory dictionaries of English. Our pipeline maps the output of a state-of-the-art dependency parser to subgraphs over concept nodes corresponding to the words of each definition. The resulting graphs have been used successfully for measuring semantic similarity (Recski and Ács, 2015) and also allows us to map virtually all English text to the 4lang representation. The full pipeline is available for download under an MIT license at http://github.com/kornai/4lang, graphs created from three major dictionaries (Longman, Collins, en.wiktionary) are also freely accessible at http://people.mokk.bme.hu/~recski/4lang/graphs. An online demo of core 4lang functionalities is available at http://4lang.hlt.bme.hu.

This paper is structured as follows: Section 2 provides a short introduction to graph-based representations of meaning, followed by an overview of the 4lang formalism and its basic principles of semantic representation in Section 3. Section 4 presents the dict_to_4lang tool, including the mapping from Stanford dependencies to 4lang configurations, and reports some figures characterizing the graphs created from each dataset. Section 5 mentions some errors typical in the output and discusses possible solutions, Section 6 presents the results of manual evaluation. Section 7 presents a method for reducing the vocabulary of newly built 4lang-graphs by replacing nodes with their definitions. Finally, Section 8 discusses some applications of the pipeline.

2. Background

Directed graphs of concepts have been used to represent the meaning of words, phrases, and utterances by several influential systems in the second half of the 20th century, including the *Semantic Memory Model* of (Quillian, 1968) or the KL-ONE system (Brachman and Levesque, 1985) and its descendants (Moser, 1983; Brachman et al., 1983). More recently, *Abstract Meaning Representation* (AMR) (Banarescu et al., 2013) was proposed as a formalism for representing meaning using directed graphs. Tools for generating AMRs from raw text have followed (Vanderwende et al., 2015; Peng et al., 2015; Pust et al., 2015), and AMRs have since been applied to a variety of NLP tasks (Pan et al., 2015; Liu et al., 2015). The 4lang theory of semantic representation (Kornai, 2010; Kornai et al., 2015), only the formalism of which can be summarized in this paper, is most similar to Quillian's model. Like the Memory Model, 4lang maps words to concepts that are defined by networks of other concepts, and allows only a very small set of relationships in such networks. 4lang differs in many aspects from AMRs, most notably by being language-independent and by limiting severely the total number of representational primitives.

The tens of thousands of graphs built by dict_to_4lang provide an important building block in the broader task of assigning 4lang representations to utterances of arbitrary size, which in turn can be used in a variety of applications in computational semantics. An early experiment applying 4lang to domain-specific understanding of natural language is presented in (Nemeskey et al., 2013), a more recent application to measuring the semantic similarity of sentences is documented in (Recski and Ács, 2015).

3. The 4lang formalism

4lang is both a formalism for representing meaning via directed graphs of concepts and the name of a manually built lexicon of such representations for ca. 2700 words [1]. A formal presentation of the system is given in (Kornai

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

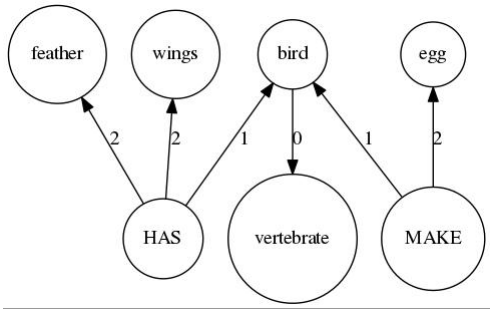

Figure 1: 4lang definition of `bird`.

et al., 2015), the theoretical principles underlying `4lang` are presented in (Kornai, 2010), we shall provide a short overview only.

`4lang` meaning representations are directed graphs of concepts with three types of edges. The most common is the 0-edge, which represents attribution (`dog` $\xrightarrow{0}$ `friendly`); the `IS_A` relation (hypernymy) (`dog` $\xrightarrow{0}$ `animal`); and unary predication (`dog` $\xrightarrow{0}$ `bark`). Edge types 1 and 2 connect binary predicates to their arguments, e.g. `cat` $\xleftarrow{1}$ `catch` $\xrightarrow{2}$ `mouse`). There are no ternary or higher arity predicates, see (Kornai, 2012). A typical definition in the `4lang` dictionary can be seen in Figure 1.

`4lang` is agnostic to parts-of-speech and voice, e.g. it makes no distinction between the words *freeze* (N), *freeze* (V), *freezing*, and *frozen*. Since attribution and (unary) predication are also treated alike, there is also no difference made between the meanings of *water freezes* and *frozen water*, both of which are represented by `water` $\xrightarrow{0}$ `freeze`.

4. Building definition graphs

The `dep_to_4lang` module implements a mapping from the output of the Stanford Dependency Parser (DeMarneffe et al., 2006) to `4lang`-subgraphs over concept nodes corresponding to words of a sentence. The `dict_to_4lang` tool extends this functionality by including parsers for three monolingual dictionaries of English – the Longman Dictionary of Contemporary English (LDOCE) (Bullon, 2003), the Collins COBUILD dictionary (Sinclair, 1987) and also database dumps of the English Wiktionary[2] – and some preprocessing steps that handle issues specific to each dataset.

To process the output of the Stanford Parser we created manually a mapping from relations to `4lang` graph configurations (presented in Table 1).

To map words to `4lang` concepts we first lemmatized them using the `hunmorph` morphological analyzer (Trón et al., 2005) and the `morphdb.en` database. We use the `ROOT` relation in the parser's output to identify the head of the definition phrase and we add a 0-edge leading to the matching concept from the headword's node. Finally we added edges to the graph based on the above mapping. The resulting graphs are the new (approximate) `4lang` definitions of each concept; an example is shown in Figure 2. Here the system correctly added edges based on *"a large wild animal that has yellow and black lines on its body"* but failed to process the remainder of the definition *"and is a member of the cat family"*. A future version of our pipeline that is still under development will also map certain combinations of dependencies, in this case the triplets *cop(member, is)* and *rcmod(animal, member)* will together trigger the edge `animal` $\xrightarrow{0}$ `member`. Finding the right representation for noun compounds such as *cat family* remains an unsolved problem, although there are plans to implement noun compound analysis in future versions of the Stanford Parser (De Marneffe and Manning, 2008).

The resulting sets of definition graphs for each dataset can be freely downloaded

[2] `https://dumps.wikimedia.org/enwiktionary/`

Dependency	Edge
amod advmod npadvmod acomp dep num prt	$w_1 \xrightarrow{0} w_2$
appos	$w_1 \xleftrightarrow{0\ 0} w_2$
nsubj csubj xsubj agent	$w_1 \xrightarrow{1} w_2$
dobj pobj nsubjpass csubjpass pcomp xcomp	$w_1 \xrightarrow{2} w_2$
poss prep_of	$w_2 \xleftarrow{1} \texttt{HAS} \xrightarrow{2} w_1$
tmod	$w_1 \xleftarrow{1} \texttt{AT} \xrightarrow{2} w_2$
prep_with	$w_1 \xleftarrow{1} \texttt{INSTRUMENT} \xrightarrow{2} w_2$
prep_without	$w_1 \xleftarrow{1} \texttt{LACK} \xrightarrow{2} w_2$
prep_P	$w_1 \xleftarrow{1} \texttt{P} \xrightarrow{2} w_2$

Table 1: Mapping from dependency relations to `4lang` subgraphs

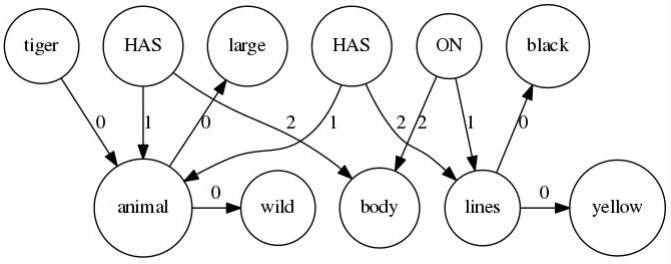

Figure 2: Definition built from: **tiger** - *'a large wild animal that has yellow and black lines on its body and is a member of the cat family'*

from `http://people.mokk.bme.hu/~recski/4lang/graphs/` as serialized python objects (`.pickle` files) that can be loaded by the `4lang` module. An interactive demo is also available under `http://4lang.hlt.bme.hu`. Table 2 shows for each dataset the total number of (non-empty) graphs and the average number of nodes in a graph.

Dict	# graphs	av. nodes
LDOCE	24 799	6.1
Collins	45 311	4.9
en.wikt	120 670	5.4

Table 2: Basic figures for each dataset

5. Issues

While the above mapping yields good results for most dictionary definitions, there are several structures that will currently result in incorrect graphs and need more sophisticated treatment than a simple mapping from dependency relations to `4lang` edges. Heads of the relations *nsubj, csubj,* etc. may be unary or binary predicates, which require different treatment in `4lang`, e.g. the relation *nsubj(eat, wombat)* should map to `wombat` $\overset{1}{\leftarrow}$ `eat` while *nsubj(smile, wombat)* warrants `wombat` $\overset{0}{\rightarrow}$ `smile`. A possible way out could be adding the latter edge for all occurences of `nsubj, csubj,` etc., claiming that the 0-relation includes all subject-predicate relations, and adding a 1-edge only in the presence of a direct object (e.g. *dobj(eat, leaf)*). This strategy would map the sentences *The wombat is eating* and *The wombat is eating a leaf* to the graphs `wombat` $\overset{0}{\rightarrow}$ `eat` and `wombat` $\overset{1}{\leftarrow}\overset{0}{\rightarrow}$ `eat` $\overset{2}{\rightarrow}$ `leaf`, respectively.

Dependencies related to quantification (*quantmod*, etc.) are not handled yet, nor are determiners or negation. Non-finite verbal modifiers of NPs (`vmod`) are also untreated, since the dependencies don't tell us if the nouns are subjects or objects of the verb in question (compare *The man climbing the tree was tall* and *The tree climbed by the man was tall*, which trigger *vmod(man, climb)* and *vmod(tree, climb)* respectively), although these cases might prove simple to disambiguate based on POS-tags in the future.

Finally, the largest number of errors are caused by incorrect parse trees, many of which are assigned to definitions that are truly ambiguous. An example is the PP attachment problem, resulting in our incorrect graph for `basement` in Figure 4, built from the Longman definition *a room or area in a building that is under the level of the ground.* Many such ambiguities are easily resolved by humans based on world knowledge (in this case e.g. that buildings with some underground rooms are more common than buildings that are entirely under the ground, if the latter can be called buildings at all), and efforts to include distributional meaning models in parsing have been reported to improve accuracy on such structures (Socher et al., 2013).

One frequent class of parse errors involve constituents modifying a coordinated phrase, which are often analysed as

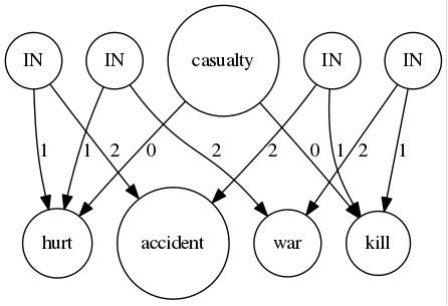

Figure 3: Definition graph built from: **casualty** - *someone who is hurt or killed in an accident or war*

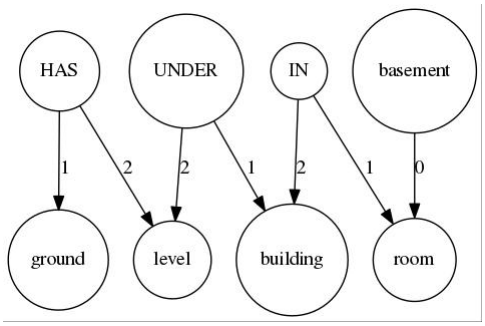

Figure 4: Incorrect definition graph built from: **basement** - *a room or area in a building that is under the level of the ground.*

modifying only one of the coordinated elements, e.g. in **casualty** - *someone who is hurt or killed in an accident or war*. We introduced a workaround to deal with these structures: in a postprocessing step edges in the `4lang` graph are copied between coordinated words (see Figure 3).

Finally, a notable error class consists of dictionary definitions that have an unusually complex phrase structure. The majority of headwords in each of our datasources are defined using a single phrase, e.g. `koala` is defined in LDOCE as *an Australian animal like a small grey bear with no tail that climbs trees and eats leaves*. In a much smaller number of cases, a full sentence containing the headword is used in definitions, e.g.:

- **playback** - *the playback of a tape that you have recorded is when you play it on a machine in order to watch or listen to it*
- **indigenous** - *indigenous people or things have always been in the place where they are, rather than being brought there from somewhere else*
- **ramshackle** - *a ramshackle building or vehicle is in bad condition and in need of repair*

Such full sentences yield a higher number of dependency relations, resulting in a denser definition graph with a higher number of erroneous edges.

6. Evaluation

To perform quantitative evaluation of our pipeline, we manually inspected a small output sample, graphs built for 20

words that were chosen randomly from the Longman Dictionary[3]. When grouping the graphs by quality we found that 11 graphs were perfect or near-perfect definitions (see e.g. Figure 5) and a further 4 were mostly accurate, with only minor details missing or an incorrect relation present in addition to the correct ones. While such a small sample obviously cannot lead us to the conclusion that 75% of graphs built by `dict_to_4lang` are of acceptable quality, these results are nevertheless promising. In a second round of evaluation we inspected all intermediate representations of the 20 definitions and grouped them based on the source of errors in the output. We found that 6 out of the 9 graphs that had errors at all were mostly affected by parser errors, while 3 were cases of the non-standard definitions discussed in Section 5.

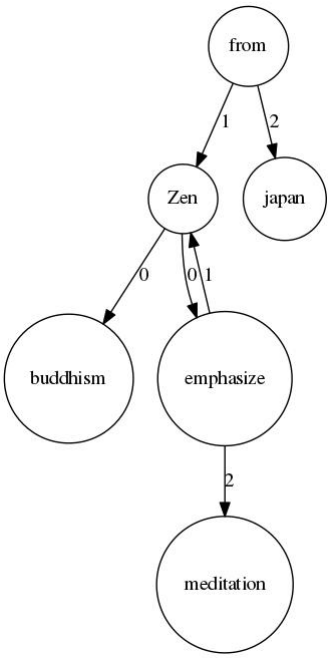

Figure 5: Graph constructed from the definition of **Zen**: *a kind of Buddhism from Japan that emphasizes meditation*

7. Expansion

The `4lang` dictionary contains by design all words of the Longman Defining Vocabulary (LDV, (Boguraev and Briscoe, 1989)). This allows us to map the words of each Longman definition to concepts that have been defined manually. This allows us to perform an *expansion* step on graphs built using `dict_to_4lang`: each node is replaced by its definition graph in the `4lang` dictionary until only those that belong to some *basic vocabulary* remain. That such vocabularies (*Feedback Vertex Sets* (FVS) of the directed graphs containing all `4lang` definitions) exist and are significantly smaller than e.g. the `4lang` dictionary itself was shown in (Kornai et al., 2015). In particular, definitions in the `4lang` dictionary can be stated using no more than 129 primitives.

[3]The 20 words in our sample, selected randomly using GNU `shuf` were the following: *aircraft, characteristic, clothesline, contrived, cypress, dandy, efface, frustrate, incandescent, khaki, kohl, lizard, nightie, preceding, residency, rock-solid, scant, transference, whatsit, Zen*

8. Applications

Since our pipeline works on any English sentence, we have also created an extension which processes running text, creates a `4lang` graph for each sentence, then merges nodes with the same label and also nodes that refer to the same entity according to the Stanford Coreference Resolution system (Lee et al., 2011). While limited by the quality of parsing, coreference resolution, and the shortcomings of our method described in Section 5, the resulting system is capable of creating a graph representation of the meaning of any English text.

The `4lang` definitions built from the Longman Dictionary using our pipeline have been used successfully in a state-of-the-art system for measuring semantic similarity of sentence pairs (Recski and Ács, 2015). This system derives sentence similarity scores from the similarity between pairs of words, and defines word similarity by measuring the overlap between `4lang` definition graphs for each word, ranking 11th out of 78 systems on the 2015 Semeval Task for Semantic Textual Similarity (Agirre et al., 2015).

9. Acknowledgements

The author wishes to thank András Kornai, Márton Makrai, Dávid Nemeskey, and two anonymous reviewers for their many useful comments on earlier versions of this paper.

10. Bibliographical References

Agirre, E., Banea, C., Cardie, C., Cer, D., Diab, M., Gonzalez-Agirre, A., Guo, W., Lopez-Gazpio, I., Maritxalar, M., Mihalcea, R., Rigau, G., Uria, L., and Wiebe, J. (2015). SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, CO, U.S.A.

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Boguraev, B. K. and Briscoe, E. J. (1989). *Computational Lexicography for Natural Language Processing*. Longman.

Brachman, R. and Levesque, H. (1985). *Readings in knowledge representation*. Morgan Kaufman Publishers Inc., Los Altos, CA.

Brachman, R. J., Fikes, R. E., and Levesque, H. J. (1983). KRYPTON: A functional approach to knowledge representation. *IEEE Computer*, 10:67–73.

Bullon, S. (2003). *Longman Dictionary of Contemporary English 4*. Longman.

De Marneffe, M.-C. and Manning, C. D., (2008). *Stanford typed dependencies manual*. Revised for the Stanford Parser v. 3.5.1 in February 2015.

DeMarneffe, M.-C., MacCartney, W., and Manning, C. (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.

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

Kornai, A. (2010). The algebra of lexical semantics. In Christian Ebert, et al., editors, *Proceedings of the 11th Mathematics of Language Workshop*, LNAI 6149, pages 174–199. Springer.

Kornai, A. (2012). Eliminating ditransitives. In Ph. de Groote et al., editors, *Revised and Selected Papers from the 15th and 16th Formal Grammar Conferences*, LNCS 7395, pages 243–261. Springer.

Lee, H., Peirsman, Y., Chang, A., Chambers, N., Surdeanu, M., and Jurafsky, D. (2011). Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task. In *Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task*, pages 28–34. Association for Computational Linguistics.

Liu, F., Flanigan, J., Thomson, S., Sadeh, N., and Smith, N. A. (2015). Toward abstractive summarization using semantic representations. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1077–1086, Denver, Colorado. Association for Computational Linguistics.

Moser, M. (1983). An overview of NIKL, the new implementation of KL-ONE. *Research in Knowledge Representation and Natural Language Understanding*, pages 7–26.

Nemeskey, D., Recski, G., Makrai, M., Zséder, A., and Kornai, A. (2013). Spreading activation in language understanding. In *Proc. CSIT 2013*, pages 140–143, Yerevan, Armenia. Springer.

Pan, X., Cassidy, T., Hermjakob, U., Ji, H., and Knight, K. (2015). Unsupervised entity linking with abstract meaning representation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1130–1139, Denver, Colorado, May–June. Association for Computational Linguistics.

Peng, X., Song, L., and Gildea, D. (2015). A synchronous hyperedge replacement grammar based approach for AMR parsing. In *Proceedings of CoNLL 2015*, page 32.

Pust, M., Hermjakob, U., Knight, K., Marcu, D., and May, J. (2015). Parsing English into Abstract Meaning Representation using syntax-based machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 1143–1154.

Quillian, M. R. (1968). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, 12:410–430.

Recski, G. and Ács, J. (2015). MathLingBudapest: Concept networks for semantic similarity. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 543–547, Denver, Colorado. Association for Computational Linguistics.

Sinclair, J. M. (1987). *Looking up: an account of the COBUILD project in lexical computing*. Collins ELT.

Socher, R., Bauer, J., Manning, C. D., and Ng, A. Y. (2013). Parsing with compositional vector grammars. In *The 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013)*.

Trón, V., Gyepesi, G., Halácsy, P., Kornai, A., Németh, L., and Varga, D. (2005). Hunmorph: open source word analysis. In Martin Jansche, editor, *Proceedings of the ACL 2005 Software Workshop*, pages 77–85. ACL, Ann Arbor.

Vanderwende, L., Menezes, A., and Quirk, C. (2015). An AMR parser for English, French, German, Spanish and Japanese and a new AMR-annotated corpus. In *Proceedings of NAACL-HLT*, pages 26–30.