Automatically Created Noun Explanations for English

Marie Stará

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
413827@mail.muni.cz

Abstract. In this paper, I comment on the automatically created explanations of word meaning for English nouns. These explanations are built using data gathered from Word Sketches created by a special Sketch Grammar.

Keywords: explanation, corpora, word sketch

1 Introduction

Following my work on automatic creation of dictionary definitions—or more precisely word meaning explanations—for Czech ([1], [2], [3]), I am modifying the method for English, so as to find out whether my approach is applicable in other languages. Hence, the purpose of this paper is to test the usability of the conversion on a smaller corpus, show the automatically created explanations of English nouns and evaluate the results.

The purpose of these explanations is to approximate the meaning of any given word by offering a set of hints (possibly) useful for understanding it.

2 Construction of Explanations

The explanations are created from the Word Sketches acquired using specially developed Sketch Grammar applied to the British National Corpus. Using Python script I take first three lemmata with the highest score for each relation and merge them in groups described below and demonstrated on the results for lemma house.

house
similar meaning can have (a/an) home, sale, garden, flat, shop, building, room
public, upper, big house
for example (a/an) home, hotel
can have/contain (a/an) garden, bedroom, room
(a/an) family, time, people can have/contain (a/an) house
is a subject of build, stand, belong

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2019, pp. 83–87, 2019.

is an object of terrace, buy, build
of (a/an) parliament, card, lord
with (a/an) garden, roof, wall

The first group of words contains hypernyms and (loose) synonyms, combinig results of five relations and data from Thesaurus (provided by Sketch Engine). These relations are: (1) *and_other* (two nouns connected by *and/or other/different/additional/further/more/such/next/similar*), (2) *WORD_is* (two nouns connected with lemma to be, possibly with other words (excluding nouns and verbs) in between), (3) *N_coord* (two nouns connected by and/or/neither-nor/either-or), and (4, 5) *hypo_hypero* (more complicated rules, basically two nouns connected either with *and (also) similar*, or *is type of*). These results are shown as a "similar meaning".

similar meaning can have (a/an) home, sale, garden, flat, shop, building, room

These similarities are followed by the list of most specific adjective modifiers of the given noun (relation *adj_modif*).

public, upper, big house

The next line of the explanation shows relation that occurs only in less than 2/3 of the test set; *for_example* (nouns connected by *(for/such) example/instance/e.g./as/like*).

for example (a/an) home, hotel

The following parts of the explanation are related to holonymy and meronymy (partitive). The fist part—"lemma can have/contain results"—is created combining relations *WORD_has* (basically two nouns connected by verb to have) and *consists of* (simply put, two nouns connected by *(can) consist/make/form/comprise/contain/include/incorporate/embody/involve/hold/cover (of)*). The second part—"results can have/contain lemma"—uses the relation *what_who_ has_WORD* (again basically two nouns connected by verb to have).

can have/contain (a/an) garden, bedroom, room
(a/an) family, time, people can have/contain (a/an) house

Other group of results is formed by verbs for which the given word is a *subject* or *object*, hence showing the results the results of relations *is_subject* and *is_object*, respectively.

is a subject of build, stand, belong
is an object of terrace, buy, build

Last two lines show results created primarily for Czech, which are also (a bit surprisingly) useful in English explanations: *of* (using the relation *gen*, two nouns connected by *of*) and *with* (sing the relation *instr*, two nouns connected by *with*).

of (a/an) parliament, card, lord
with (a/an) garden, roof, wall

3 Evaluation of Explanations

I evaluated the explanations on a set of 70 nouns; for the sake of comparison, I used the translation of my test set for Czech (hence I deleted words with multi-word translation as is, e.g. *fish breeding ground* (trdliště). Six of these words never occurred in the corpus; two words have frequency lower than twenty. Ten words have frequency in between 20 and 100. The highest frequency in the test set is 67 826 (*song*). It is apparent from these data that the test set contains words of various frequencies. Another distinction between the tested expressions is the presumed difficulty of creating the explanation. (It is rather straightforward to explain meaning of e.g. *a dog* (an animal which barks) or *a house* (a building for living). Explaining what a is e.g. *nothingness* (absence of anything?) or *laughter* (loud expression of happiness?) is supposedly more challenging, especially should the explanation be reasonably short and specific. This distinction is most visible in between abstract and concrete nouns.

The test set contains words with more meanings, homonyms (*(a) lead—to lead*, and synonyms (*couch*, *sofa*). Some words were picked ad hoc to ensure the test set is sufficiently differentiated.

3.1 Examples

As mentioned above, the test set contains lemmata with zero frequency in the corpus. Three words are interpreted as a different part of speech (*an expose, (a) lead* are recognised/appearing in the corpus only as verbs, *a mammoth* as an adjective).

There are also other words, for which the Word Sketches do not yield sufficient data, e.g. *mamluk* (frequency 35). Apart from this extreme case, there are other words without enough good data, e.g. *excavator*, where only the first two lines (and arguably the *is_object* relation) contain sensible result.

mamluk
is an object of exist

excavator
similar meaning can have (a/an) digger, tractor, shaker, extractor, servo, pride mini, rotary, bulk excavator

for example (a/an) bow, hall, town
can have/contain (a/an) coin, finance, right
(a/an) other can have/contain (a/an) excavator
is a subject of indemnify, assign, exploit
is an object of alert, prompt, protect
of (a/an) heart

There are not only quite good results (see *house* above) and rather bad ones, for the most part one explanation contain ballanced amount of good and bad (and okay-ish) results, e.g. *bed*.

bed
similar meaning can have (a/an) breakfast, border, table, room, door, wall
double, twin, unmade bed
for example (a/an) doctor, child
(a/an) price, kitchen, room can have/contain (a/an) bed
is a subject of stare, knit, sleep
is an object of share, wet, strip
of (a/an) rose, lettuce, nail
with (a/an) flu, someone, sheet

An often appearing problem is badly recognised part of speech in the data. This can be seen e.g. in the explanation of *oak*, where there are nouns recognised as verbs (e.g. *pine* in all occurrences in coordination *oak and pine*).

3.2 Evaluation

Generally, the best results are the adjective modifiers, followed by the word with similar meaning and the of-relation. Surprisingly, the explanations are more reliable when the given lemma is an object, not a subject of a verb.

Significantly fewer results are found for the has/contains relations, as well as the with-relation. These result might change when a bigger corpus is used. The least reliable results are the lists of examples, where there is a lot of noise.

As hinted in 3.1, one of the reasons the explanations contain irrelevant (or simply wrong) data is wrongly tagged tokens. Nevertheless, it is apparent the results show the chosen approach is applicable and can be used with minor editing.

4 Conclusions

With the corpus data containing mistakes in part of speech tags, it is quite difficult to automatically create sufficient explanation for any given word. The results are, nevertheless, encouraging as bigger data generally lead to better results.

Acknowledgements. This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin infrastructure LM2015071.

References

1. Stará, M. Automatická tvorba definic z korpusu. Masters thesis, Masaryk University (2019)
2. Stará, M. Automatically Created Noun Definitions for Czech. Proceedings of the Twelfth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2018 (2018)
3. Stará, M., Kovář, V. Options for Automatic Creation of Dictionary Definitions from Corpora. Tenth Workshop on Recent Advances in Slavonic Natural Language Processing, RASLAN 2016 (2016)