UDC: 004.934.2 https://doi.org/10.15407/jai2024.03.094

M. Klymenko
Institute of Artificial Intelligence Problems of the Ministry of Education and Science of Ukraine and the National Academy of Sciences of Ukraine, Ukraine
40, Akademika Glushkova Ave., Kyiv, 03680
nik@ipai.net.ua
https://orcid.org/0000-0003-4433-6641

RELATION MEASUREMENT BETWEEN SEMANTIC FIELDS BY METRIC APPROACH

Abstract. The article considers a numerical research of approach for semantic metric between lexical units calculation. Received a set of statistical characteristics of the lexicographic semantic trees. Simplified representation of tree as a semantic field is proposed and operations for relation measurement between fields is described. This approach can be used for explainable language model creation for natural language processing tasks.
Keywords: semantic field, lexicographic semantic tree, attributes of semantic relation.

I. Introduction

At present natural language processing consists of a large number of tasks. Large language models development for machine learning techniques made it possible to automate analytical tasks such as text translation, generation of thematic texts, content summarization and correction of errors in them. The weaknesses of current applied methods based on language models are the lack of ontological knowledge and context-dependent natural language expressions semantics.

The extracting of semantic information is designed to improve the accuracy of terms usage, translation and will also contribute to research for elements and models development of general artificial intelligence [1].

In this paper, numerical research of approach for semantic metric between lexical units calculation is carried out. Based on it results relation measurement method for semantic fields is proposed.

II. Relation measurement method

A. Description of semantic field construction.

For current numerical research we follow the approach [2], which formulates a machine-friendly expression for semantic fields describing. Word meaning is represented by a set of semantically grouped words. It is useful to perform unsupervised extraction of grouping characteristics by processing large text corpus. Instead of this, we propose to use dictionary definitions as the source of pregrouped semes by referenced word.

Thus we represent the meaning of a word by the geometric sum of the semes contained in its dictionary definition:

$$w_0 = \frac{1}{\sqrt{n_0}} \sum_{i_1}^{n_0} \mathrm{w}_i, \qquad (1)$$

where n_0 is the number of semes w_i, contributing to the meaning of initial word w_0.

By claiming that meanings of the words at every semantic level are linear combinations of the meanings of the words at the preceding level, (1) is generalized to describe each w_i seme, we can formalize general w_{i_k} seme:

$$w_{i_k} = \frac{1}{\sqrt{n_{i_k}}} \sum_{i_{k+1}}^{n_{i_k}} \mathrm{w}_{i_k i_{k+1}}. \qquad (2)$$

Received recursive semantical decomposition can be naturally represented by tree structure. With the tree depth increasing, the appearance of lexicographic hyperchains and hypercycles becomes inevitable [3] as described on Fig. 1.

In example (Fig. 1) we stop expanding tree on the nodes where hyperchains was finded (“human” and “existence” semes).

Semantic field in used approach can be obtained by (2) to represent vector in multidimensional semantical space. Since the construction of the complete semantical space requires processing of the vast majority of lexical units, we are using a simplified semantic field representation. It is composed as a algebraic sum of weighted by (2) tree nodes.

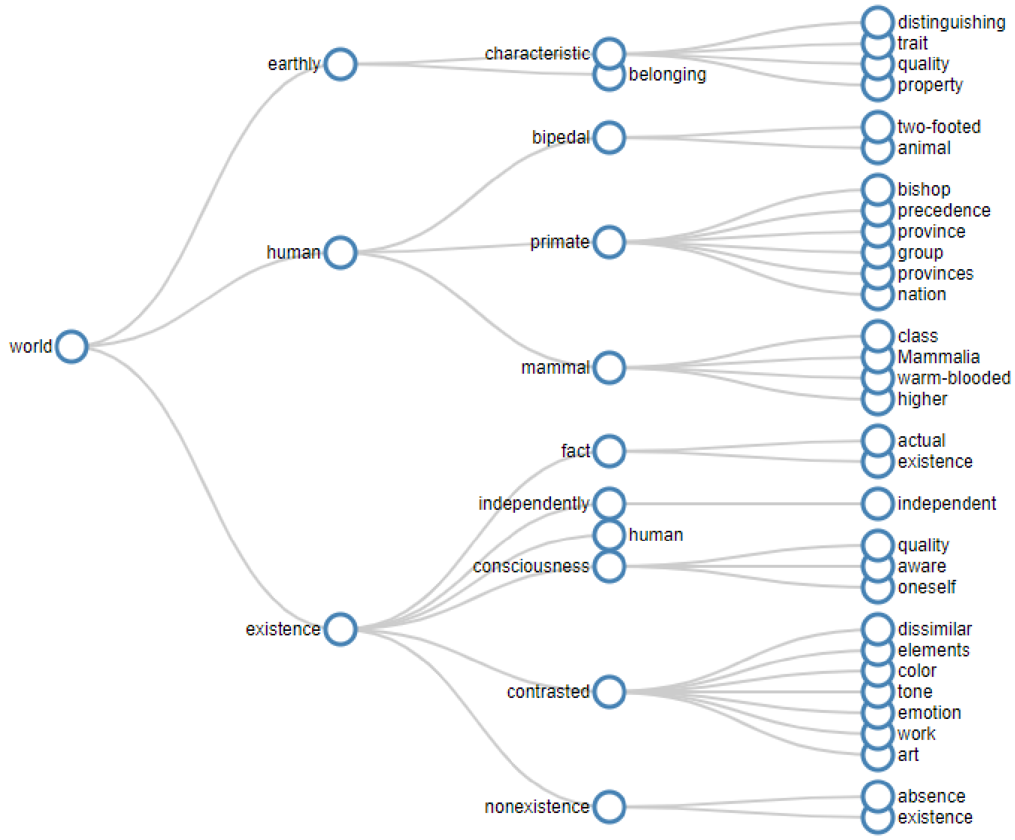

Fig. 1. Example of 3-level semantic tree builded up from “world”

$$w_{i_k} = \frac{1}{\sqrt{n_{i_k}}} \frac{1}{a_{i_k}} \sum_{i_{k+1}}^{n_{i_k}} a_{i_k i_{k+1}} \mathrm{w}_{i_k i_{k+1}}. \qquad (3)$$

In this way, the contribution of each seme to the semantic characteristic of the word is determined. As tree build up is limited practically by its depth, we should take into account hypercycles and hyperchains that were dropped out on lower levels for optimizing reasons. Their subtrees should be multiplied by corresponding coefficient of the seme.

B. An approach for relation measurment between fields.

As a result of semantic field build up (3) the meaning of the word represented by set:

$$w = \{s_1, s_2 \dots, s_{n-1}, s_n\}, \qquad (4)$$

where s_n is weight of named seme. Assuming we have received w_1 and w_2 semantic fields with n_1 and n_2 number of seme weights respectively. We propose to compare fields in order to be able to determine different lexical semantic relations such as synonymy, antonymy, meronymy. Also, field comparison can provide characteristics for relation description between semantic fields. Using operations on sets we can obtain a number of basic quantitative indicators:

- subsets of common and different semes, their absolute number and related to comparative sets;
- weighted sum of subsets;
- ranked position of subsets related to ordered by weight comparative sets.

Based on this indicators relation between semantic fields can be measured as:

$$rel(w_1, w_2) = \frac{r(w_1 \cap w_2)}{r(w_1 \oplus w_2)}, \qquad (5)$$

where *r()* describes sum of subset ranks. Relation (5) tends to 1 and can’t be calculated only in cases when $w_1 \in w_2$ or vice versa. The use of rank instead of weight is suggested because of the possible cumulative weight advantage of low-ranked semes. Similarly, the difference of semantic fields can be calculated as a antonymy descriptor:

$$diff(w_1, w_2) = \frac{\sum w_1 \oplus w_2}{\sum w_1 \cup w_2}. \quad (6)$$

To research the expediency of described approach we perform several numerical simulations.

III. Numerical research

To automate semantic field build up we use popular "Dictionary by Merriam-Webster" available on the Internet and includes some 470,000 entries [4]. Python app was created to parse online pages with definitions in multithread mode. Due to peculiarities of dictionary markup and approach conditions, some specialized processing rules were introduced:

- added list of stopwords, which are filtered out from set of semes [5];
- among the multitude of definitions of a polysemous word, priority is given to the noun;
- additionally definition of a word without ending is checked (for the possibility of finding synonyms);
- lexicographic hypercycles and hyperchains check on prior tree levels to speed-up the processing.

The tree is bypassed "in width" for hyperchains consideration. During research 200 semantic fields of widely used English words was builded. The depth of tree was limited to 10, however single field was performed down to 15 level with more than 25 thousand semes in it.

General statistical data about received fields (Table 1) shows main trends of seme number by tree levels. The quantitative distribution of values is given, taking into account the confidence interval $p < 0,05$.

Table I. Statistical data on 200 semantic trees

Depth level	**Average number of added semes**	**Average number of duplicated semes**
1	11 ± 54%	0
2	52 ± 63%	1 ± 100%
3	144 ± 83%	13 ± 88%
4	340 ± 62%	81 ± 79%
5	521 ± 64%	247 ± 73%
6	992 ± 58%	812 ± 76%
7	1836 ± 61%	1108 ± 69%
8	3115 ± 56%	3244 ± 62%
9	5792 ± 49%	7802 ± 43%
10	10105 ± 38%	16621 ± 40%

According to received statistic of added semes we can see an assured descent of increase rate from more than 3 times at upper levels to below 2 times at the deepest levels. This is complemented by a stable decrease of values distribution. With it we notice dramatically growth of amount of duplicated semes. These trends correspond to expectations and confirm the rationality of limiting the depth of tree construction. Therefore further numerical evaluation performed limited by 4-level depth trees.

Gathered semantic fields were intersected pairwise to evaluate relation characteristics between them and look at numerical dependence on semantic relations such as synonymy and antonymy. Table 2 shows top-20 by weight common semes of "downpour" and "rain" semantic fields. While most weighted of them can be associated with definition of selected words, other semes look like randomly gathered in intersected list with slightly meaning relation to words.

Table II. Most weighted common semes of "downpour" and "rain" semantic fields

Word	**Weight**
process	0,1420
action	0,1260
relating	0,0336
range	0,0275
conditions	0,0149
thought	0,0146
series	0,0138
power	0,0138
place	0,0116
higher	0,0108
open	0,0107
placing	0,0097
arranging	0,0097
circumstances	0,0097

objects	0,0097
numbers	0,0097
vertebrate	0,0084
extended	0,0084
order	0,0077
instance	0,0064

Another semantic fields intersection is showed on Table 3. Selected words aren't synonyms, but they have deep semantic relation, which is confirmed by top-20 semes. Some of them such as water, condition, air, atmosphere, temperature, surface and earth are remarkable descriptors of semantic relation between words meanings. Other semes can't be associated with neither of two words instead.

Table III. Most weighted common semes of "snow" and "winter" semantic fields

Word	Weight
quality	0,0334
water	0,0208
process	0,0203
person	0,0157
body	0,0136
action	0,0122
instance	0,0111
character	0,0099
material	0,0086
relating	0,0085
condition	0,0075
air	0,0074
visible	0,0073
atmosphere	0,0068
temperature	0,0065
marked	0,0062
surface	0,0061
earth	0,006
characteristic	0,0059
fact	0,0057

Table 4 shows another similar example of two semantically connected fields intersection. Result coincides in general: up to 7 of top-20 semes describe meaning of words and their semantical relation. It should be noticed, appropriate semes have not biggest weights in separate fields and in intersection.

Table IV. Most weighted common semes of "rain" and "snow" semantic fields

Word	Weight
quality	0,0334
water	0,0208
process	0,0203
person	0,0157
body	0,0136
small	0,0134
action	0,0122
instance	0,0111
property	0,0103
particles	0,0099
matter	0,0087
material	0,0086
relating	0,0085
liquid	0,0081
condition	0,0075
air	0,0074
visible	0,0073
atmosphere	0,0068
tasteless	0,0068
marked	0,0062

A similar situation persists in comparison with other semantically related pairs of words. We extend experiment in order to prove a trend is noticed. Table 5 shows semes intersection of related words as the seasons of the year, so similarity in definitions can be achieved on upper levels of their semantic trees. Amount of appropriate semes slightly increases comparing with pairwise intersections, but there are still a lot of semes barely semantically connected with definitions.

Table V. Most weighted common semes of "winter", "summer", "spring" and "autumn" semantic fields

Word	Weight
number	0,0314
process	0,0233
hemisphere	0,0172
time	0,0169
relating	0,0151
quality	0,0129
period	0,0106
marked	0,0070
definite	0,0043

position	0,0042
brought	0,0040
natural	0,0038
good	0,0037
group	0,0036
based	0,0035
order	0,0029
strength	0,0026
atmosphere	0,0068
tasteless	0,0068
marked	0,0062

As expected, we receive some irrelevant to main definition semes in separate semantic field due to insufficient context in seme selection, which should be made more selectively. However, their presence in the intersections of a large number of fields is unacceptable because they impair the quality of semantic connections building and their attributes extracting. We intersect all 200 gathered sematic fields to describe the direct dependence of the common semes number increase with the deepening of the semantic tree. Results are shown on Table 6.

Table VI. Common semes number of 200 semantic trees by teir level

Depth level	**Common semes number**
1-6	0
7	8
8	20
9	53
10	191

For intersections of large number of fields it is normal to use more than 5-level depth trees. For cleaning pairwise intersections from poor relevant semes a field of background semes is proposed. As an approach for gathering of such semes could be ranking of them in multiple subfields, collected from synonyms fields intersections, where inappropriate semes can be easily defined.

The collected set of background field semes then sould be used to filter out semantic fields. We assume to exclude from the field (or subset) semes which rank is equal or lower than in background field. This assumption is corresponded with current numerical investigation but should be examined further.

Conclusion

The numerical simulations of metric approach [2] was performed. They show the rationality of limiting the depth of tree construction. The question can be discussed: weight of seme at the deep level for the parent word meaning and reasonable depth limitation.

Empirically was showed that there are almost 2% (of average number in field) common semes in builded up to 10 level depth semantic fields. Research should be continued on more representative array of vocabulary words. However, on the basis of received data it is proposed to build background semantic field. Might be useful to throw out of consideration such seme subset taking into account their ranking and weight in target semantic fields.

More complex comparisons should be processed on phrases, expressions, sentences and texts. Further work in this direction will allow to verify current results and to examine collected semantic information in practical natural language processing tasks.

References

1. Shevchenko, A.I. “Natural Human Intelligence - The Object of Research for Artificial Intelligence Creation” International Scientific and Technical Conference on Computer Sciences and Information Technologies, 2019, 1, pp. XXVI–XXIX, DOI: 10.1109/STC-CSIT.2019.8929799.
2. Vakulenko, Maksym. “From Semantic Metrics to Semantic Fields.” 2021 IEEE 16th International Conference on Computer Sciences and Information Technologies (CSIT) (2021): pp.44-47, DOI: 10.1109/CSIT52700.2021.9648675.
3. Shyrokov V. “System semantics of explanatory dictionaries” Cognitive Studies, Nov. 2015, pp. 95-106, DOI: 10.11649/cs.2012.007.
4. Dictionary by Merriam-Webster, 2022. URL: https://www.merriamwebster.com/.
5. NLTK's list of english stopwords, 2010. URL: https://gist.github.com/sebleier/554280.

The article has been sent to the editors 18.03.24.
After processing 25.04.24.
Submitted for printing 30.09.24