Hybrid Knowledge-based and Data-driven Text Similarity Estimation based on Fuzzy Sets, Word Embeddings, and the OdeNet Ontology

Tim vor der Brück

School of Computer Science and
Information Technology
Lucerne University of Applied Sciences and Arts
Rotkreuz, Switzerland
e-mail: tim.vorderbrueck@hslu.ch

Michael Kaufmann

School of Computer Science and
Information Technology
Lucerne University of Applied Sciences and Arts
Rotkreuz, Switzerland
e-mail: m.kaufmann@hslu.ch

***Abstract*—Estimating the semantic similarity between texts is important for a wide range of application scenarios in natural language processing. With the increasing availability of large text corpora, data-driven approaches such as Word2Vec have become quite successful. In contrast, semantic methods, which employ manually designed knowledge bases such as ontologies, have lost some of their former popularity. However, manually designed expert knowledge can still be a valuable resource, since it can be leveraged to boost the performance of data-driven approaches. In this paper, we introduce a novel hybrid similarity estimate based on fuzzy sets that exploits both word embeddings and a lexical ontology. As ontology, we use OdeNet, a freely available resource developed by the Darmstadt University of Applied Sciences. Our application scenario is targeted marketing, in which we aim to match people to the best fitting marketing target group based on short German text snippets. The evaluation showed that the use of an ontology did indeed improve the overall result in comparison with a baseline data-driven estimate.**

Keywords–OdeNet; fuzzy sets; targeted marketing; histogram equalization.

I. Introduction

Note that this paper is an extended version of [1]. In comparison with the original conference paper, we updated some of the linguistic resources (stop word list, lemmatization, and OdeNet ontology), conducted additional experiments and gave a more detailed evaluation. In particular, we evaluated three additional coefficients for our ontology-based similarity estimate, namely the Sørensen-Dice coefficient (henceforth, the *Dice coefficient*), the overlap coefficient, and pointwise mutual information. Furthermore, we investigated the distribution of the gold standard annotations and determined milieu-wise precision, recall, and F1-scores for the most accurate similarity estimate.

The approach presented here was developed in cooperation with a marketing company with the goal of facilitating market segmentation, which is one of the key tasks of a marketer. Usually, market segmentation is accomplished by clustering demographic variables, geographic variables, psychographic variables, and behaviors [2]. In this paper, we will describe an alternative approach based on unsupervised natural language processing. In particular, our business partner operates a commercial youth platform for the Swiss market, where registered members receive access to third-party offers such as discounts and special events (e.g., concerts or castings). Several hundred online contests per year, which are sponsored by other firms, are launched over this platform, and an increasing number of them require members to write short free-text snippets (e.g., to elaborate on a perfect holiday at a destination of their choice in case of a contest sponsored by a travel agency). Based on the results of a broad survey, the platform provider's marketers assume six target groups (called *milieus*) exist among the platform members. For each milieu (with the exception of the default milieu *Special Groups*) a keyword list was manually created to describe its main characteristics. To trigger marketing campaigns, an algorithm has been developed that automatically assigns each contest answer to the most likely target group: we propose the youth milieu as the best match for a contest answer, for which the estimated semantic similarity between the associated keyword list and user text snippet is maximal. For the estimation of text relatedness, we devised a novel semantic similarity estimate based on a combination of word embeddings and OdeNet (*Offenes deutsches Wordnet - open German wordnet*), where the latter is a freely available lexical ontology recently developed by the Darmstadt University of Applied Sciences.

The remainder of this paper is organized as follows: In the next section, we survey some of the related work in the area of semantic text similarity estimation. Our proposed methodology is described in Section III. Section IV introduces the OdeNet ontology and compares it with GermaNet. In Section V, we investigate the way similarity estimates can be combined that exhibit very different probability distributions. The obtained evaluation results are given in Section VI and discussed in Section VII. Finally, we conclude the paper in Section VIII with an overview of the accomplished results and possible future work.

II. Related Work

There is a multitude of existing approaches to estimating text similarity by means of ontologies. Liu and Wang [3] match each word of a text to a concept in an ontology and derive a vector representation for it consisting of its weighted one-hot-encoded hypernyms, hyponyms, and the matched concept itself, where the weights are specified beforehand and they assume the maximum value of 1 for the latter. An entire document can then be represented by the centroid vector

of all words in the documents. As usual, the comparison with other documents can be accomplished by applying the cosine measure on the centroids. In contrast to Liu and Wang, Mabotuwana et al. [4] disregard the hyponyms for constructing the word vectors and set the weight of a hypernym to the reciprocal of the number of nodes on the shortest path in the ontology from the matched concept to the hypernym. A downside of this method is that simple path length count is quite unreliable in capturing semantic similarity, which is a finding of Resnik [5]. Therefore, the latter introduced information content (IC), which is the negative logarithm of the occurrence probability of a word and aims to compensate for differences of semantic similarities between nodes of taxonomy edges. The IC constitutes also the basis for several novel semantic similarity measures introduced by Lastra Díaz et al. [6], [7]. Mingxuan Liu and Xinghua Fan [8] propose enriching texts with semantically related words (hyponyms/hypernyms) to improve the categorization of short Chinese texts, which is the approach, we want to follow here. However, in contrast to Mingxuan Liu and Xinghua Fan, we will not represent the words occurring in the texts by ordinary sets but instead by fuzzy sets, that allow us to incorporate word vectors in our similarity score. The approach using fuzzy sets has the additional advantage that very general hypernyms or overly specific hyponyms, which are not really related to the input texts anymore and possibly introduce noise, can be downvoted.

All the state-of-the-art methods described so far return a single scalar value as a similarity estimator. however, Oleshshuk and Pedersen's approach derives a similarity vector, which represents the semantic similarities on different abstraction levels of the ontology as estimated by the Jaccard index [9].

An alternative approach to estimate semantic similarity is the use of word embeddings. These embeddings are determined beforehand on a very large corpus typically using either the skip-gram or the continuous bag-of-words variant of the Word2Vec model [10]. The skip-gram method aims to predict the textual surroundings of a given word by means of an artificial neural network. The influential weights of the one-hot-encoded input word to the nodes of the hidden layer constitute the embedding vector. For the so-called *continuous bag-of-words* method, it is just the opposite, i.e., the center word is predicted by the words in its surrounding. Alternatives to Word2Vec are GloVe [11], which is based on aggregated global word co-occurrence statistics and Explicit Semantic Analysis (ESA) [12], in which each word is represented by the column vector in the tf-idf matrix over Wikipedia. The idea of Word2Vec can be transferred to the level of sentences as well. In particular, the Skip-Thought vector model [13] derives a vector representation of the current sentence by predicting the surrounding sentences. An alternative to Skip-Thought vectors are Bert Sentence Embeddings that are based on a transformer architecture [14]. If vector space representations of the documents are established, a similarity estimate can then be obtained by applying the cosine measure on the embeddings centroids of the two documents to compare.

There is some prior work to devise similarity estimates combining ontologies and word embeddings. Faruqui et al.'s [15] approach aims to retrofit the embedding vectors in such a way that related words with respect to the employed ontology have preferably similar vector representations. Goikoetxea et al. [16] generate random walks on WordNet to extract sequences of concepts. These sequences are then fed into the ordinary Word2Vec to create (ontology) embeddings vectors. They evaluated several possibilities to combine such vectors with word embeddings, such as averaging or concatenating them. A downside of this approach in comparison with our proposed estimate is that at least 1 million of such random walks must be generated to obtain sufficiently reliable results. Therefore, the required format conversion, which needs to be repeated for every change in the ontology, is quite time-consuming.

III. Proposed Method

A straightforward and simple method to estimate the similarity between two texts is applying the Jaccard index on their bag-of-words representations [17, p. 299]. This coefficient is given as:

$$jacc(A,B) := \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where A (B) is the set of words of the first (second) text. In this scenario, the first text is the snippet entered by the user and the second text is the keyword description of the youth milieu.

An alternative to the Jaccard index is the Dice coefficient [17], which is defined as follows:

$$DSC(A,B) := \frac{2|A \cap B|}{|A| + |B|} \tag{2}$$

One can define distance measures for these two coefficients, which are called the Jaccard distance and the Dice distance, respectively, by subtracting them from 1. In contrast to the Jaccard distance, the Dice distance does not satisfy the triangle inequality [18, p. 29] and is therefore not a proper distance metric. Note that the Dice coefficient and Jaccard index can be transformed into each other by the following formulas:

$$\begin{aligned} jacc(A,B) &= DSC(A,B)/(2 - DSC(A,B)) \\ DSC(A,B) &= 2\,jacc(A,B)/(1 + jacc(A,B)) \end{aligned} \tag{3}$$

Furthermore, we consider the overlap coefficient, which is given by [17, p. 299]:

$$overlap(A,B) := \frac{A \cap B}{\min\{|A|, |B|\}} \tag{4}$$

It assumes the maximum value of 1 if either one of the input sets is a subset of the other.

While these coefficients work reasonably well for long texts, they usually fails for short text snippets since in this case, it is very likely that all overlaps are caused by very common words (typical stop words), which are actually irrelevant for estimating text similarity. One possibility to increase the number of overlaps is to extend the two texts by means of an ontology [8], i.e., adding to a text the words from the ontology that are semantically close (hence reachable by a short path) to the words of that text. In particular, we decided to add all synonyms, hypernyms, and direct hyponyms of all words appearing in the investigated text. Hereby we follow the hypothesis of Rada et al. [19], which states that taxonomic relations are sufficient to capture semantic similarity between ontology concepts. Note that hyponyms and hypernyms may not be uniquely defined since a single word can occur in

several synsets. In principle, two possibilities to deal with arise in this situation:

1) Use hyponyms / hypernyms of all possible synsets for the expansion
2) Employ Word Sense Disambiguation to select only the synset that corresponds to the indented meaning of the word. The drawback of this approach is that the Word Sense Disambiguation might choose the incorrect synset, especially with short text snippets, which can result in missing overlaps and therefore inexact similarity estimates.

Currently, we use possibility 1 but consider possibility 2 for a future version of our approach.

The two sets used in the coefficients stated above (Jaccard, Dice, and Overlap) are crisp, which means that all words are treated alike. However, the words that are newly induced by the ontology are probably less reliable for capturing the semantics of the text than the original words are. Furthermore, not all of the newly introduced words are equally relevant. However, our current model cannot capture those relationships. Therefore, we extend our set representation to allow for fuzziness (i.e., we employ fuzzy sets instead of conventional crisp sets).

For conventional sets, the decision of whether an element belongs to this set is always binary (i.e., it can uniquely be decided whether an element belongs to a set or not). This is different from a fuzzy set, where the membership of an element can be partial. In particular, each fuzzy set is assigned a real-valued function $\mu : X \rightarrow [0, 1]$ (X: all potential elements of our set) assuming values in the interval [0,1] and specifying the degree of membership for all elements. If this membership function only assumed the values 0 or 1, the fuzzy set would actually be equivalent to a conventional set.

Set union and intersection are also defined in terms of fuzzy sets, namely in the following way:

$$\begin{aligned} \mu_{A \cap B} &= \min\{\mu_A, \mu_B\} \\ \mu_{A \cup B} &= \max\{\mu_A, \mu_B\} \end{aligned} \tag{5}$$

The cardinality of a fuzzy set is defined as the total sum over all membership values:

$$|F| := \sum_{x \in X} \mu_F(x)$$

With intersection, union, and fuzzy set cardinality, all three coefficients described above (Jaccard, Dice, and Overlap) can be defined for fuzzy sets analogously to ordinary sets.

In addition to these coefficients, we also employ pointwise mutual information, which is defined as:

$$pmi(A, B) := lb\left(\frac{P(A \cap B)}{P(A)P(B)}\right) \tag{6}$$

where $A \cap B$ denotes the Fuzzy set intersection between A and B. The probability of a fuzzy event represented in the form of a fuzzy set E is given by $|E|/n$ [20], where n denotes the number of elements in the fuzzy set, in this case all lemmas of the German language possibly occurring in one of the texts. Note that the cardinality of E is defined as the sum of all fuzzy membership values and is therefore different from n.

To avoid dealing with negative infinity values of the pointwise mutual information, which occur if the sets A and B are disjoint, we follow the approach of [21] and clip all values less than -2.0. To combine the pointwise mutual information with the Word2Vec-based similarity estimate, we linearly scale all its values into the interval $[0, 1]$.

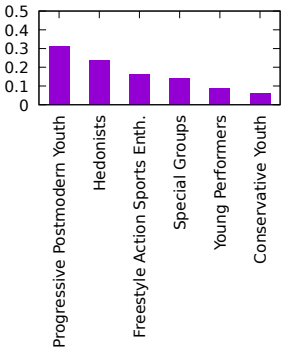


Figure 1. Distribution of youth milieus over the three contests.

What remains is to define the fuzzy membership function. Let $Cent(A)$ be the word embeddings centroid of our original words. We then define the membership function μ as follows:

$$\mu(w) := (\max\{0, \cos(\angle(Cent(A), Emb(w)))\})^i \tag{7}$$

where $Emb(w)$ is the embedding vector of a word w and the use of the maximum operator prevents the membership value from being complex. The exponent i allows us to adjust the influence of the word embeddings gradually. Full influence is obtained by setting i to 1. In contrast, the influence diminishes if i is set to 0.

Our similarity estimate is then used to assign user responds from several online contests in form of short text snippets to the best fitting youth milieu out of *Progressive Postmodern Youth* (people primarily interested in culture and arts), *Young Performers* (people striving for a high salary with a strong affinity to luxury goods), *Freestyle Action Sports Enthusiasts*, *Hedonists* (rather poorly educated people who enjoy partying and disco music), and *Conservative Youth* (traditional people with a strong concern for security). A sixth milieu called *Special Groups* comprises all those who cannot be assigned to one of the upper five milieus. The distribution of the 6 milieus over the three considered contests is given in Figure 1 as a histogram. This figure shows that the milieus are quite unevenly distributed with the most frequent milieu *Progressive Postmodern Youth* appearing around five times more often than the rarest one (*Conservative Youth*).

For each milieu (with the exception of *Special Groups*) a keyword list was manually created to describe its main characteristics (see Table I). To trigger marketing campaigns, an algorithm has been developed that automatically assigns each contest answer to the most likely target group: we propose the youth milieu as the best match for a contest answer, for which the estimated semantic similarity between the associated keyword list and user respond is maximal. In case the highest similarity estimate falls below the 10 percent quantile for the distribution of highest estimates, the *Special Groups* milieu is selected.

TABLE I. KEYWORD LISTS DESCRIBING THE YOUTH MILIEUS.

Youth milieu	Keywords
Progressive Postmodern Youth	clothing, music, art, freedom, culture, educated
Young Performers	rich, elite, luxury, luxurious
Freestyle Action Sports Enthusiasts	Sports, Fitness, Music
Hedonists	poor, communication, self-fulfilment, entertainment, party, music, disco
Conservative Youth	conservation of value, conservativism, citizenship, Switzerland

TABLE II. EXAMPLE USER ANSWER FOR THE TRAVEL DESTINATION CONTEST (TRANSLATED INTO ENGLISH).

Choice	Country	Snippet
1	Jordan	Ride through the desert and marvel at Petra during sunrise before the arrival of tourist buses
2	Cook Island	Snorkelling with whale sharks and relax
3	USA	Experiencing an awesome week at the Burning Man Festival

The ontology we employ for our similarity estimate is OdeNet, which is a freely available lexical resource recently developed by the Darmstadt University of Applied Sciences that will be explained in more detail in the next section.

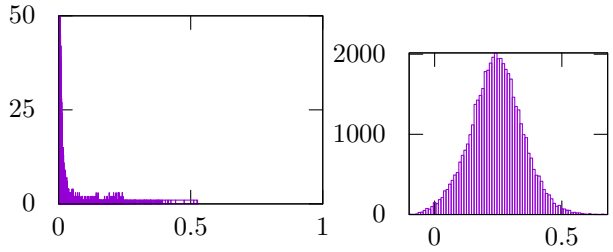
IV. ODENET ONTOLOGY

Freely available machine-readable lexical ontologies for German are rather sparse. On the one hand, there are websites such as Wiktionary and Open-Thesaurus, which are targeted at human users. Much effort would need to be spent to bring the associated resources to a form that can be exploited efficiently by a computer. On the other hand, there is GermaNet [22], which is suitable both for human users as well as for automated processing. However, GermaNet is not a free resource. While it may be freely used in purely academic projects, as soon as industry partners are involved, the academic license is no longer eligible and the project partners have to sign a commercial license agreement.

The lexical ontology of OdeNet [23] is devised to fill this gap. It has been compiled automatically from the Open-Thesaurus synonym lexicon (https://www.openthesaurus.de/), the Princeton WordNet of English [24], and the Open Multilingual WordNet English [25]. Afterwards, it was manually error-checked and applied to comprehensive revisions. Similar to WordNet, semantic concepts are represented by synsets, which are interconnected by linguistic and semantic relations such as hyponymy, hypernymy, meronymy, holonymy, and antonymy. In total, it currently contains around 120 000 lexical entries and 36 000 synsets. The entire resource is available as an XML file obtainable at Github [26]. We found OdeNet to be very easy to use and well-designed.

V. COMBINING SIMILARITY SCORES

Besides our ontology based measure, we implemented several other measures such as ESA, the cosine of word embedding centroids, Skip-Thought vectors, etc. Usually, a stronger and more reliable similarity estimate can be obtained by combining measures. One possibility for that is majority vote, i.e., suggesting the class that most of the measures



(a) Ontology-based estimate (Jaccard Index).

(b) Cosine of embeddings centroids.

Figure 2. Histograms of similarity estimates.

suggest. One drawback of majority vote is that the individual measures should be of comparable performance and that we need at least three of them. Furthermore, a majority vote only returns a decision for one of the classes but no (numerical) score. However, we actually need such a score to determine the 10 percent quantile (cf. previous section). An alternative to a majority vote is a weighted average. Albeit, there is again an obstacle. While all our semantic similarity estimates assume values between 0 and 1 (Note that the cosine of word embeddings centroids can assume (usually small) negative values as well.), their distributions can be quite different (see Figure 2). Considering this case, we would like to combine the cosine of word embedding centroids and our ontology based similarity measure by a weighted sum. The first type of estimate is normally distributed and covers almost the entire value range. However, although in principle our ontology based similarity estimate can reach the value of 1, most of its values are located inside the interval [0,0.1]. To make both estimates comparable with each other, we are conducting a histogram equalization for them prior to their combination. Such an equalization levels out the relative occurrence frequencies of estimate intervals, so that the resulting values are approximately uniformly distributed. This is accomplished by transforming the similarity estimates using the cumulative probability distribution function cdf. Formally, an estimate s is mapped to the value $cdf(s)$. One downside of our method is that the resulting similarity estimate is probably biased. However, in our scenario, we are not so much interested in the actual value of our estimate but instead focusing mainly on the correct ranking of target groups. Thus, the modification of the estimate's probability distribution is unproblematic. The combined estimate sim is formally given as:

$$sim := w \cdot cdf(sim_{odenet}) + (1 - w) \cdot cdf(sim_{w2v}) \quad (8)$$

where

- w: in the influencing weight of the OdeNet similarity based estimate, the default value is 0.5
- sim_{odenet}: the score obtained by the OdeNet similarity estimate (Jaccard, Dice, Overlap, or Pointwise Mutual Information over fuzzy sets)
- sim_{w2v}: cosine of the angle between the Word2Vec embeddings centroids of user text snippet and youth milieu keyword description

TABLE III. OBTAINED ACCURACY VALUES FOR SEVERAL SIMILARITY ESTIMATES. ODENET+EMB.: LINEAR COMBINATION OF OUR ONTOLOGY BASED MEASURE WITH COSINE OF WORD EMBEDDINGS CENTROIDS. RW=RANDOM WALK BASED METHOD PROPOSED BY GOIKOETXEA ET AL., STV=SKIP-THOUGHT VECTORS, JC=JACCARD INDEX, OL=OVERLAP COEFFICIENT, PMI=POINTWISE MUTUAL INFORMATION, HE=HISTOGRAM EQUALIZATION [16]

Method	Contest 1	2	3	Total
Random	0.172	0.149	0.197	0.172
Jaccard	0.150	0.194	0.045	0.142
W2V	0.348	**0.328**	0.227	0.330
ESA	0.357	0.254	**0.288**	0.335
RW	0.281	0.149	0.273	0.263
Bert	0.109	0.149	0.136	0.118
STV	0.162	0.284	0.273	0.191
Emb.+JC	0.266	0.313	0.227	0.267
Emb.+JC (HE)	0.347	**0.328**	0.227	0.330
OdeNet(JC,crisp)	0.367	0.194	0.273	0.333
OdeNet(JC)	0.309	0.224	0.212	0.286
OdeNet(JC)+Emb.	**0.380**	0.269	0.273	**0.352**
OdeNet+Emb.+Mero	0.372	0.254	0.273	0.345
OdeNet(OL)+Emb.	0.370	0.209	**0.288**	0.339
OdeNet(Dice)+Emb.	0.372	0.254	0.273	0.345
OdeNet(PMI)+Emb.	0.370	0.224	**0.288**	0.341

TABLE IV. MINIMUM AND MAXIMUM AVERAGE INTER-ANNOTATOR AGREEMENTS (COHEN'S KAPPA).

Method	Contest 1	2	3
Min kappa	0.123	0.295/0.030	0.110/0.101
Max. kappa	0.178	0.345/0.149	0.114/0.209
# Annotated entries	1543	100	100

VI. EVALUATION

For evaluation, we selected three online contests (language: German), where people elaborated on their favorite travel destination (contest 1, see Table II for an example), speculated about potential experiences with a pair of fancy sneakers (contest 2) and explained why they emotionally prefer a certain product out of four available candidates. In a bid to provide a gold standard, three professional marketers from different youth marketing companies annotated independently the best matching youth milieus for every contest answer. We determined for each annotator individually his/her average inter-annotator agreement with the others (Cohen's kappa). The minimum and maximum of these average agreement values are given in Table IV. Since for contests 2 and 3, some of the annotators considered only the first 50 entries (last 50 entries respectively), we specified min/max average kappa values for both parts.

Before automatically distributing the texts to the youth milieus, we applied on them a linguistic preprocessing consisting of tokenization, stop word filtering, lemmatization, and compound analysis. The latter was used to determine the base form of each word, which was added as an additional token. Next to our own similarity estimates, we evaluated several baseline methods, in particular random assignments, Jaccard, ESA, the ontology-based approach of Goikoetxea et al. [16], cosine of word embedding centroids, Skip-Thought vectors, and Bert embeddings. The accuracy values given in Table III are obtained by comparing automated assignments with the majority vote of the assignments conducted by our human annotators. Since the keyword lists used to describe the characteristics of the youth milieus typically consist of nouns (in the German language capitalized) and the user contest answers might contain many adjectives and verbs as well, which do not match nouns very well in the Word2Vec vector

TABLE V. CORPUS SIZES MEASURED BY NUMBER OF WORDS.

Corpus	# Words
German Wikipedia	651 880 623
Frankfurter Rundschau	34 325 073
News journal *20 Minutes*	8 629 955

TABLE VI. PRECISION, RECALL AND F1-SCORE OBTAINED FOR THE THE YOUTH MILIEUS USING THE ONTOLOGY-BASED ESTIMATE (JACCARD-INDEX).

Milieu	Precision	Recall	F1-score
Special Groups	0.548	0.453	0.496
Freestyle Action Sports Enthusiasts	0.374	0.506	0.430
Hedonists	0.287	0.565	0.380
Progressive Postmodern Youth	0.507	0.211	0.298
Young Performers	0.091	0.064	0.075
Conservative Youth	0.200	0.032	0.056
All	0.335	0.305	0.289

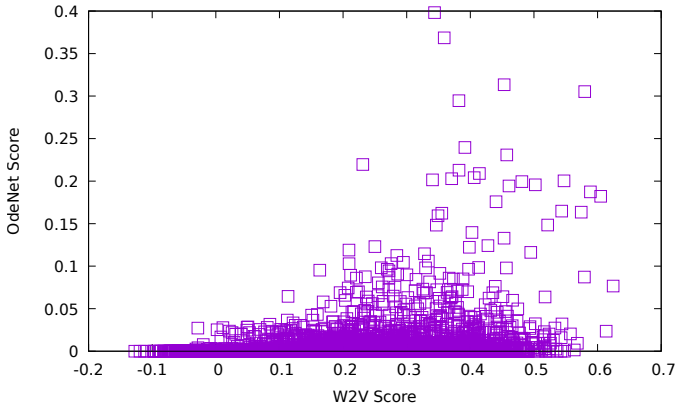


Figure 3. Scatterplot: Word2Vec (W2V) Embeddings Score vs OdeNet Score.

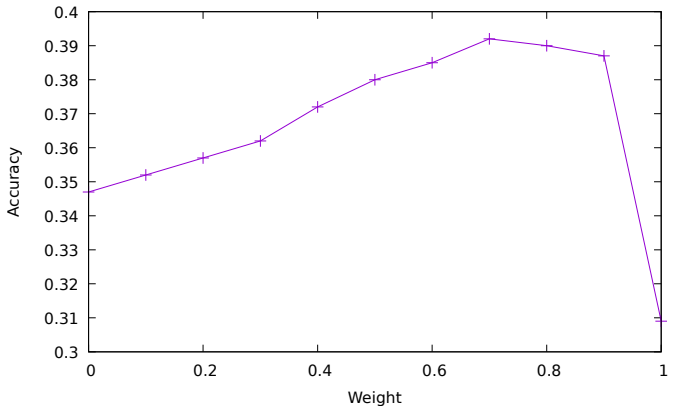


Figure 4. Accuracy of combined Word2Vec Embeddings and OdeNet score with respect to the influencing OdeNet score weight w.

representation, we actually conduct two comparisons for the Word2Vec centroids based on similarity estimate, one with the unchanged user contest answers and one by capitalizing every word beforehand. The final similarity estimate is then given as the maximum value of both individual estimates. For our proposed ontology based similarity estimate, we use the parameter settings $i = 0.5$ and weights of linear combination: 0.5, which performed best in several experiments with varying parameter values. Setting i to 0.5 seems to us to be a good compromise between considering only the ontology structure ($i = 0$) and fully weighting the word embedding vectors ($i = 1$).

In total, the following methods are evaluated:

1) Random: Just randomly assign one of the youth milieus to a text snippet
2) Jaccard: Estimate the semantic similarity of a text snippet and youth milieu keywords by applying the Jaccard index directly to their bag of words representations
3) Word2Vec: Estimate the semantic text similarity by applying the cosine measure to the word embedding centroids
4) ESA: Estimate the semantic text similarity by applying the cosine measure to the ESA embedding centroids
5) RW: Similarity estimate based on random walks over an ontology (here: OdeNet, in the original paper: WordNet) as proposed by Goikoetxea et al.
6) Bert: Estimate the semantic text similarity based on the centroids of Bert embeddings
7) STV: Estimate the semantic text similarity based on the centroids of Skip-Thought vectors
8) Emb.+JC: Averaging estimates of methods 2 and 3
9) Emb.+JC (HE): The same as above but additionally conducting a histogram equalization.
10) OdeNet (JC,crisp): OdeNet based similarity measure using the Jaccard with exponent i set to 0 which results in crisp (non-fuzzy) sets
11) OdeNet (JC): OdeNet based similarity estimate employing Jaccard index with exponent i set to 0.5
12) OdeNet (JC)+Emb.: Averaging estimates of methods 3 and 11
13) OdeNet (JC)+Emb+Mero: Averaging estimates of methods 3 and 11, where in method 11, the lemmas are expanded not only by hyponyms but also by meronyms
14) OdeNet (OL)+Emb: Similar to method 11 but instead of Jaccard the overlap index is used
15) OdeNet (Dice)+Emb: Similar to method 11 but instead of Jaccard the Dice index is used
16) OdeNet (PMI)+Emb: Similar to method 11 but instead of Jaccard Pointwise Mutual Information is used

The Word2Vec word embeddings were trained on the German Wikipedia (dump originating from 20 February 2017) merged with a Frankfurter Rundschau newspaper corpus and 34 249 articles of the news journal *20 minutes*, where the latter is targeted to the Swiss market and freely available at many Swiss train stations (see Table V for a comparison of corpus sizes). By employing articles from *20 minutes*, we want to ensure the reliability of word vectors for certain Switzerland specific expressions such as *Velo* or *Glace*, which are underrepresented in the German Wikipedia and the Frankfurter Rundschau corpus.

The accuracy of the combined OdeNet / Word2Vec embedding score with respect to the weight of the OdeNet score is given in Figure 4. This diagram shows that the maximum accuracy value is obtained at a weight of $w = 0.8$, which demonstrates that a rather large OdeNet weight is required for obtaining a high accuracy. This fact seems a bit surprising, since the standalone OdeNet similarity estimate performs much poorer than its Word2Vec embedding counterpart.

Furthermore, we give the F1-score of the combined OdeNet / W2V embedding similarity estimate for the individual youth milieus in Table VI. The highest F1-score is obtained for the *Special Groups* youth milieu, which insinuates that oftentimes the appropriate youth milieu is not expressed by the contest participants in the text snippets. The second-best detectable milieu is *Freestyle Action Sports Enthusiasts*, which is caused by the fact that in the first and largest contest containing elaborations of possible dream holidays, the participants frequently mention sports activities such as the surfing or snorkeling that they plan to conduct.

Finally, the scatter plot of the OdeNet (Jaccard) vs Word2Vec embedding similarity estimate is specified in Figure 3. This plot demonstrates that the relationship between both estimates is highly nonlinear and the Ontology-based estimate frequently scores text pairs rather low that assume a high Word2Vec Embeddings estimate value.

VII. Discussion

The evaluation shows that although our ontology based method lags behind the cosine of Word2Vec centroids in terms of accuracy, their linear combination performs considerably better than both of the methods alone. Furthermore, it outperforms both its crisp counterpart (exponent i:=0), the approach of Goikoetxea et al. if applied to OdeNet, used with 100 million random walk restarts, and combined with Word2Vec word embeddings by vector concatenation (RW in Table III) and also two deep learning based approaches (Skip-Thought vectors and Bert embeddings [14]). The rather low accuracy of both deep learning approaches (Skip-Thought and Bert) is caused by the fact that the words of the keyword lists describing the youth milieus are arbitrarily ordered and therefore these lists can not be captured sufficiently well by a language model trained on ordinary texts like Wikipedia. In further experiments, we could show that especially Bert embeddings are very vulnerable to ungrammatical input. For instance, a simple stop word filtering degrades its performance already considerably.

Remarkable is the low performance of our approach on contest 2. Further analysis revealed that in several cases the correct youth milieu in this contest was indicated by the only word that was either a town name (“Basel”) or a rather rare noun that is not contained in OdeNet, which demonstrates that the given ontology is indeed very useful for estimating semantic similarity.

Note that the OdeNet ontology is still under active development and contains several gaps in the semantic relations. For instance, it comprises no hyponyms of *sports*, which makes it difficult to correctly assign people to the *Freestyle Action Sports Enthusiasts* target group. Another downside is that OdeNet contains no inflected forms so far. Thus, we have to employ a lemmatizer in order to identify hyponyms and hypernyms for such word forms.

VIII. Conclusion and Future Work

We presented a similarity estimate based both on word embeddings and OdeNet ontology. In contrast to most state-of-the-art methods, it can directly employ the given ontology format. Time consuming format conversions into vectors or matrices are not necessary, which simplifies its usage significantly. Additionally, by using fuzzy sets, hypernyms/hyponyms introduced by the ontology that are too general/specific and therefore not really related to the input texts any more, can be downvoted. The application scenario is targeted marketing, in which we aim to match people to the best fitting marketing target group based on short German text snippets. The evaluation showed that the obtained accuracy of a baseline method considerably increases if combined by a linear combination with our ontology based estimate. In general, this estimate attains a good performance, if the ontology contains the key terms relvant for the application scenario. As future work we want to further investigate hybrid data-driven and knowledge-based semantic similarity estimates. In particular, we plan to employ additional semantic relations besides hypernyms, hyponyms, synonyms, and meronyms such as holonyms or antonyms. Furthermore, all the model parameters are currently manually specified. It would be preferable to determine them automatically through the use of grid search or more sophisticated Artificial Intelligence methods such as Bayesian search [27]. Finally, we want to experiment with other types of hierarchically ordered lexical resources, which are not necessarily ontologies, such as the Wikipedia category taxonomy.

Acknowledgment

We thank Jaywalker GmbH as well as Jaywalker Digital AG for their support regarding this publication and especially for annotating the contest data with the best-fitting youth milieus. This research has been funded by the FMSquare Stiftung, an international foundation for the promotion of fuzzy management methods.

References

[1] T. vor der Brück, "Estimating semantic similarity for targeted marketing based on fuzzy sets and the odenet ontology," in International Conference on Advances in Semantic Processing, 2018.

[2] M. Lynn, "Segmenting and targeting your market: Strategies and limitations," Cornell University, Tech. Rep., 2011, online: http://scholorship.sha.cornell.edu/articles/243 [retrieved: 09/2018].

[3] H. Liu and P. Wang, "Assessing text semantic similarity using ontology," Journal of Software, vol. 9, 2014.

[4] T. Mabotuwana, M. C. Lee, and E. V. Cohen-Solal, "An ontology-based similarity measure for biomedical data - application to radiology reports," Journal of Biomedial Informatics, vol. 46, 2013.

[5] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), 1995.

[6] J. J. Lastra-Díaz and A. García-Serrano, "A novel family of IC-based similarity measures with a detailed experimental survey on WordNet," Engineering Applications of Artificial Intelligence, vol. 46, 2015, pp. 140–153.

[7] J. J. Lastra-Díaz, A. García-Serrano, M. Batet, M. Fernández, and F. Chirigati, "HESML: A scalable ontology-based semantic similarity measures library with a set of reproducably experiments and a replication dataset," Information Systems, vol. 66, 2017, pp. 97–118.

[8] M. Liu and X. Fan, "A method for Chinese short text classification considering effective feature expansion," Internation Journal of Advanced Research in Artificial Intelligence, vol. 1, no. 1, 2012.

[9] V. Oleschchuk and A. Pedersen, "Ontology based semantic similarity comparison of documents," in Proceedings of the 14th International Workshop on Database and Expert Systems Applications (DEXA), 2003.

[10] T. Mikolov, I. Sutskever, C. Ilya, G. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in Proceedings of the Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, Nevada, 2013, pp. 3111–3119.

[11] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP), Doha, Katar, 2014.

[12] E. Gabrilovic and S. Markovitch, "Wikipedia-based semantic interpretation for natural language processing," Journal of Artificial Intelligence Research, vol. 34, 2009.

[13] R. Kiros et al., "Skip-Thought vectors," in Proceedings of the Conference on Neural Information Processing Systems (NIPS), Montréal, Canada, 2015.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019.

[15] M. Faruqui et al., "Retrofitting word vectors to semantic lexicons," in Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2015.

[16] J. Goikoetxea, E. Agirre, and A. Soroa, "Single or multiple? Combining word representations independently learned from text and WordNet," in Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, Arizona USA, 2016.

[17] C. D. Manning and H. Schütze, Foundations of Statistical Natural Language Processing. MIT Press, 1999.

[18] K. Latha, Experiment and Evaluation in Information Retrieval Models. Boca Raton, Florida: CRC Press, 2018.

[19] R. Rada, H. Mili, E. Bicknell, and M. Blettner, "Development and application of a metric on semantic nets," IEEE Transactions on Systems, Man, and Cybernetics, vol. 19, no. 1, 1989, pp. 17–30.

[20] O. Pavlačka1 and P. Rotterová, "Probability of fuzzy events," in 32nd International Conference on Mathematical Methods in Economics, 2014, pp. 760–765.

[21] A. Salle and A. Villavicencio, "Why so down? the role of negative (and positive) pointwise mutual information in distributional semantics," CoRR, vol. abs/1908.06941, 2019. [Online]. Available: http://arxiv.org/abs/1908.06941

[22] B. Hamp and H. Feldweg, "GermaNet - a lexical-semantic net for German," in Proceedings of the ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, 1997.

[23] M. Siegel and F. Bond, "OdeNet: Compiling a German WordNet from other resources," in Proceedings of the 12th Global WordNet Conference, 2021, pp. 192–198.

[24] C. Fellbaum, Ed., WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press. Cambridge, Massachusetts: MIT Press, 1998.

[25] F. Bond and R. Foster, "Linking and extending an open multilingual wordnet," in Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 1352–1362.

[26] M. Siegel et al., "OdeNet," last access: 11/22/2021. [Online]. Available: https://github.com/hdaSprachtechnologie/odenet

[27] J. Snoek, H. Larochelle, and R. P. Adams, "Practical Bayesian optimization of machine learning algorithms," in Proceedings of the Conference on Neural Information Processing Systems (NIPS), 2012, pp. 2951–2959.