Estimating Semantic Similarity for Targeted Marketing based on Fuzzy Sets and the Odenet Ontology

Tim vor der Brück

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

Abstract—**Estimating the semantic similarity between texts is of vital importance for a wide range of application scenarios in natural language processing. With the increasing availability of large text corpora, data-driven approaches like Word2Vec became quite successful. In contrast, semantic methods, which employ manually designed knowledge bases like ontologies lost some of their former popularity. However, manually designed knowledge can still be a valuable resource, since it can be leveraged to boost the performance of data-driven approaches. We introduce in this paper 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 recently 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

Market segmentation is one of the key tasks of a marketer. Usually, it is accomplished by clustering over demographic variables, geographic variables, psychographic variables and behaviors [1]. 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 get access to third-party offers such as discounts and special events like concerts or castings. Actually, several hundred online contests per year are launched over this platform sponsored by other firms, an increasing number of them require the 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 different target groups (called *milieus*) being present 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. For triggering 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 best match for a contest answer, for which the estimated semantic similarity between the associated keyword list and user answer is maximal. For the estimation of text relatedness, we devised a novel semantic similarity estimate based on a combination of word embeddings and Odenet, where the latter is a freely available lexical ontology recently developed by the Darmstadt University of Applied Sciences.

There is a multitude of existing approaches to estimate text similarity by means of ontologies. Liu and Wang [2] 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 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. [3] disregard the hyponyms for constructing the word vectors and set the weight of a hypernym to the inverse of the number of nodes on the shortest path in the ontology from the matched concept to this 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 [4]. Therefore, he introduced the so-called 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. [5], [6]. Mingxuan Liu and Xinghua Fan [7] propose to enrich texts with semantically related words (hypernyms) to improve the categorization of short Chinese texts, which is the approach, we want to follow here. But, 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, which allows us to incorporate word vectors in our similarity score. All the methods described so far return a single scalar value as similarity estimator. The approach of Oleshshuk and Pedersen however, derives a similarity vector, which represents the semantic similarities on different abstraction levels of the ontology as estimated by the Jaccard index [8].

An alternative method 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 [9]. 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 [10], which is based on aggregated global word co-occurrence statistics and the Explicit Semantic Analysis (ESA) [11], 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 so-called Skip-Thought Vector model [12] derives a vector representation of the current sentence by predicting the surrounding sentences. Again, a similarity estimate can be obtained by applying the cosine measure on the embeddings centroids of the two documents to compare. There is some former work to devise similarity estimates combining ontologies and word embeddings. The approach of Faruqui et al. [13] 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. [14] 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 like averaging or concatenating them. A downside of this approach in comparison with our proposed estimate is that at least one million of such random walks must be generated to obtain sufficiently reliable results. So, the required format conversion, which needs to be repeated for every change in the ontology, is quite time-consuming.

The remainder of this paper is organized as follows: Our proposed methodology is described in Section II. Section III introduces the Odenet ontology and compares it with GermaNet. In Section IV we investigate, how similarity estimates can be combined that exhibit very different probability distributions. The evaluation is contained in Section V. Finally, we conclude the paper in Section VI with an overview of the accomplished results and possible future work.

II. Proposed Methodology

A straight-forward and simple method to estimate the similarity between two texts is applying the Jaccard index on their bag of words representations [15]. 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. While this approach works reasonably well for long texts, it 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 [7], i.e., adding the words from the ontology to a text 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 the direct hyponyms of all words appearing in the investigated text. Hereby we follow the hypothesis of Rada et al. [16], 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, there are two possibilities to deal with this situation:

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

Currently, we use possibility one but consider possibility two for a future version of our approach.

The two sets used in the Jaccard index 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. 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 whether an element belongs to this set is always crisp, i.e., it can uniquely be decided if an element belongs to this 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 $\mu : X \to [0,1]$ (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:

$$\begin{aligned}\mu_{A\cap B} &= \min\{\mu_A,\mu_B\}\\ \mu_{A\cup B} &= \max\{\mu_A,\mu_B\}\end{aligned} \tag{2}$$

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

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

By transferring our method to fuzzy sets, the applied similarity measure, the Jaccard index, stays unchanged. The only difference is that we compare fuzzy sets with each other and not any more conventional sets. What remains is to define the membership function. Let $Cent(A)$ be the word embeddings centroid of our original words. We then define the membership function μ as follows:
$\mu(w) \quad := \quad (\max\{0, cos(\angle(Cent(A), Emb(w)))\})^i$ 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 gradually adjust the influence of the word embeddings. Full influence is obtained by setting i to one. In contrast, the influence diminishes if i is set to zero.

Our similarity estimate is then used to assign user answers of several online contests to the best fitting youth milieu, which

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

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

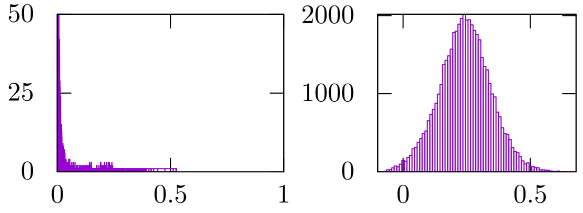
are *progressive postmodern youth* (people primarily interested in culture and arts), *young performers* (people striving for a high salary with a strong affinity to luxury goods), *freestyle action sportsmen*, *hedonists* (rather poorly educated people who enjoy partying and disco music) and *conservative 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. For each milieu (with the exception of *special groups*) a keyword list was manually created to describe its main characteristics. For triggering 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 best match for a contest answer, for which the estimated semantic similarity between the associated keyword list and user answer 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.

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 and will be explained in more detail in the next section.

III. ODENET ONTOLOGY

Freely available machine-readable lexical ontologies for German are rather sparse. On the one hand, there are websites like Wiktionary and Open-Thesaurus, which are targeted at human users. A lot of effort would have to be spent to bring the associated resources in a form that can be efficiently exploited by a computer. On the other hand, there is GermaNet [17], which is suitable both for human users as well as for automated processing. However, GermaNet is no 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 Odenet [18][19] is devised to fill this gap. It has been automatically compiled from the Open-Thesaurus, Wiktionary, and the Open Multilingual WordNet English. 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 like hyponymy, hypernymy, meronymy, holonymy and antonymy. In total, it currently contains 120 012 lexical entries and 36 192 synsets. The entire resource is available as an XML file, which can be obtained at Github [20]. We found Odenet very easy to use and well-designed.



(a) Histogram for ontology-based estimate.

(b) Histogram for cosine of embeddings centroids.

Figure 1. Histograms of similarity estimates.

IV. COMBINING SIMILARITY SCORES

Besides our ontology based measure, we implemented a whole bunch of other measures like ESA, cosine of word embeddings 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 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 zero and one (Note that the cosine of word embeddings centroids can assume (usually small) negative values as well.), their distributions can be quite different (see Figure 1). Consider the case, we would like to combine cosine of word embeddings 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 to 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 their 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 focus mainly on the correct ranking of target groups. Thus, the modification of the estimate's probability distribution is unproblematic.

V. EVALUATION

For evaluation, we selected three online contests (language: German), where people elaborated on their favorite travel destination (contest 1, see Table I for an example), speculated about potential experiences with a pair of fancy sneakers

TABLE II. 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. [14]

Method	Contest			
	1	2	3	Total
Random	0.167	0.167	0.167	0.167
ESA	0.357	0.254	**0.288**	0.335
Word2Vec Centroids	0.347	**0.328**	0.227	0.330
Skip-Thought Vectors	0.162	0.284	0.273	0.191
Odenet	0.308	0.224	0.227	0.288
Odenet+Emb.	**0.377**	0.239	0.273	**0.347**
Odenet (crisp)+Emb.	0.374	0.224	0.273	0.343
Odenet+Emb.+Mero.	0.375	0.239	0.273	0.345
RW	0.281	0.149	0.273	0.263

TABLE III. 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

(contest 2) and explained why they emotionally prefer a certain product out of four available candidates. In 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 III. Since for contests 2 and 3, some of the annotators annotated 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, lemmatization, and compound analysis. The latter was used to determine the base form of each word, which was added as additional token. Next to our own similarity estimates, we evaluated several baseline methods, in particular ESA, cosine of word embeddings centroids, Skip-Though-Vectors, and random assignments. The accuracy values given in table Table II are obtained by comparing the automated assignment 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 a lot of adjectives and verbs as well, which do not match very well to nouns in the Word2Vec vector representation, we actually conduct two comparisons for the Word2Vec centroids based 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 as a good compromise between considering only the ontology structure ($i = 0$) and fully weighting the word embedding vectors ($i = 1$). Furthermore, we evaluated enriching the input texts with meronyms in addition to taxonomic relations, which slightly decreased the obtained accuracy (Odenet+Emb.+Mero. in Table II).

TABLE IV. 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

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 various Swiss train stations (see Table IV 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 like *Velo* or *Glace*, which are underrepresented in the German Wikipedia and the Frankfurter Rundschau corpus.

The evaluation shows that although our ontology based method lags behind 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) and 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 II). Quite striking is the poor performance of our approach on contest 2. Further analysis revealed that in several cases the correct youth milieu in this contest was indicated by only one word that was either a town name ("Basel") or a rather rare noun that are not contained in Odenet.

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 sportsman* 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. However, the German model shipped with this lemmatizer is of rather mediocre quality. Therefore, we are currently building a suitable dataset to retrain the lemmatizer.

VI. CONCLUSION AND FUTURE WORK

We presented a similarity estimate based both on word embeddings and the Odenet ontology. In contrast to most state-of-the-art methods, it can directly employ the given ontology format. Time consuming format conversions are not necessary, which simplifies its usage significantly. 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. As future work we want to employ additional semantic relations besides hypernyms, hyponyms, synonyms and meronyms like holonyms or antonyms. Furthermore, all the model parameters are currently manually specified. It would be preferable to determine them automatically by the use of grid search or more sophisticated Artificial Intelligence methods like Bayesian search [21]. Finally, we want to experiment with other types of hierarchically ordered lexical resources, which are not necessarily ontologies, like the Wikipedia category taxonomy.

ACKNOWLEDGEMENT

Hereby we thank the Jaywalker GmbH as well as the Jaywalker Digital AG for their support regarding this publication and especially for annotating the contest data with the best-fitting youth milieus.

REFERENCES

[1] 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].

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] R. Kiros et al., “Skip-thought vectors,” in Proceedings of the Conference on Neural Information Processing Systems (NIPS), Montréal, Canada, 2015.

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

[14] 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.

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

[16] 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.

[17] 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.

[18] M. Siegel, “Talk: Odenet - a German contribution to the multilingual WordNet initiative (Odenet - ein deutscher Beitrag zur Multilingual Open WordNet Initiative),” 2017.

[19] ——, “Odenet,” Linguistic Issues in Language Technology - LiLT (submitted), 2018.

[20] M. Siegel et al., “Odenet,” last access: 11/12/2018. [Online]. Available: https://github.com/hdaSprachtechnologie/odenet

[21] 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.