Social Semantics And Its Evaluation By Means Of Semantic Relatedness And Open Topic Models

Ulli Waltinger
Text Techolgogy
Bielefeld University
33602 Bielefeld, Germany
ulli_marc.waltinger@uni-bielefeld.de

Alexander Mehler
Text Technology
Bielefeld University
33602 Bielefeld, Germany
alexander.mehler@uni-bielefeld.de

***Abstract*—This paper presents an approach using social semantics for the task of topic labelling by means of Open Topic Models. Our approach utilizes a social ontology to create an alignment of documents within a social network. Comprised category information is used to compute a topic generalization. We propose a feature-frequency-based method for measuring semantic relatedness which is needed in order to reduce the number of document features for the task of topic labelling. This method is evaluated against multiple human judgement experiments comprising two languages and three different resources. Overall the results show that social ontologies provide a rich source of terminological knowledge. The performance of the semantic relatedness measure with correlation values of up to .77 are quite promising. Results on the topic labelling experiment show, with an accuracy of up to .79, that our approach can be a valuable method for various NLP applications.**

***Keywords*-social semantics; semantic relatedness; open topic models; text classification; topic identification;**

I. Introduction

In this paper we consider the problem of topic identification on *Open Topic Models* (OTM). That is, we are not heading towards a clustering of a document collection but labelling individual documents with the best fitting topic names obtained from a social ontology. In this context social ontologies are used as a source of terminological knowledge providing a large-scale but most importantly a flexible knowledge system in building OTM. OTM are topic-related models in which content categories are not assigned in advance but change over time – contributed by the open community. Content categories themselves are predefined by the constantly growing social ontology itself. Our approach utilizes such a social ontology by the alignment of documents within a social network comprising category information trails. Therefore we treat the task of topic identification as a problem of ontology alignment. Doing this, we identify the documents of a collection that are most closely related for a given text fragment. We then use this article-category information to conduct a topic generalization. Our approach can be subdivided into three consecutive steps. Firstly, we build, using a social network, two vector representations (Wiki Vectors) comprising article and category concepts – described in Section III-A. Secondly, we extract the most informative lexemes within a document by proposing a text representation by means of lexical chains (Section III-D). This is done in order to reduce the complexity within the topic labelling task. Tracking and connecting semantically related tokens in a text (lexical chaining) goes along with the task of measuring word relatedness. In this paper we propose a measure for semantic relatedness on the basis of distributional feature properties – Wiki Vectors – within a social network (Section 4). Thirdly, we compute thematic labels proposing a topic generalization technique using category trails within the Wikipedia taxonomy (Section III-B). In Section IV-B we present the results of our evaluation (Section IV). Experiments on semantic relatedness are based on four different datasets and compared to various state-of-the-art approaches comprising three different resources and two languages. The task of topic identification by means of OTM is evaluated using two manually built corpora comprising 20 topics in 2000 documents. We show that our approach can be a valuable method for various NLP-applications.

II. Related Work

In recent years various approaches have been proposed to the problem of automatic modelling associations between words or text fragments. However, algorithms and results differ depending on resources and the way the experiment was set up. In general we can subdivide these approaches on the basis of their resources into three different groups: distributional, lexical-semantic net driven and Wikipedia-based methods. Distributional similarity can be defined in measures establishing relatedness on direct co-occurrence in text (1^{st} order) – e.g. frequency information of co-occurrences [1], [2], on bigrams [3], on information-based sequence distance [4], on Google page-counts (*Normalized Google Distance*) [5], [6] – or on comparing the similarity of contexts in which two terms occur (2^{nd} order). Here the *Latent Semantic Analysis* (LSA) [7] has obtained particular attention, due to its success in a large variety of tasks involv-

Table I
EXAMPLE WSR SCORES FOR DIFFERENT DOMAINS

Word	Word	Relatedness Score
Google	Sergei Brin	.784
Microsoft	Sergei Brin	.645
Microsoft	Bill Gates	.875
Yahoo	Bill Gates	.525
Federal Bureau of Investigation	FBI	.970
Central Intelligence Agency	FBI	.618
CDU (german party)	Angela Merkel	.756
SPD (german party)	Angela Merkel	.633
Angela chancellor	Merkel	.952
winter	snow	.798
summer	snow	.515

ing semantic processing. Most recently, *Semantic Vectors* [8] promise to perform as successfully as techniques like LSA, but unlike them, Semantic Vectors do not rely on complex procedures such as *Singular Value Decomposition* (SVD). Using a lexical-semantic net like Princeton *WordNet* [9], *EURO-WordNet* [10] or its German counterpart *GermaNet*, [11] numerous measures have been proposed in the past ([12], [13], [14], [15], [16]). See I. Cramer (2009) [17] for an overview of the performance of lexical-semantic net related measures. The methods mostly use a hyponym-tree induced from a given word net. With regard to Wikipedia-based semantic relatedness computation, numerous approaches have been proposed. These methods mainly focus either on the hyperlink structure [18], the vector space model (VSM) and/or on category concepts for graph related measures [19], [20]. E. Gabrilovich and S. Markovitch (2007) [21] proposed a method called *Explicit Semantic Analysis*, which represents term similarity by a high-dimensional space of article concepts derived from Wikipedia. Our work follows their approach in terms of building a reduced vector representation from a social network. Hence, computing semantic relatedness differs to the methods stated above. Since most of the results of the previous approaches were reported on different human judged datasets we will evaluate our method on the three most widely used setup scenarios. Most approaches in topic identification focus either on topic clustering techniques [22] by clustering keywords using different notions of a similarity measure [23], [24] or by an automatic text categorization [25] scenario using a small set of given categories. In this context, our approach utilizes over 55,000 different categories as topic labels and combines both keyword extraction as a type of text representation and categorization by means of topic labelling. Therefore the domain of our approach meets rather the task of ontology alignment [26], [27] inducing social ontologies by means of the Wikipedia dataset as an automatic text categorization. See Figure 1 for an overview of the proposed ontology alignment (Section III-A) and topic generalization (Section III-B) technique.

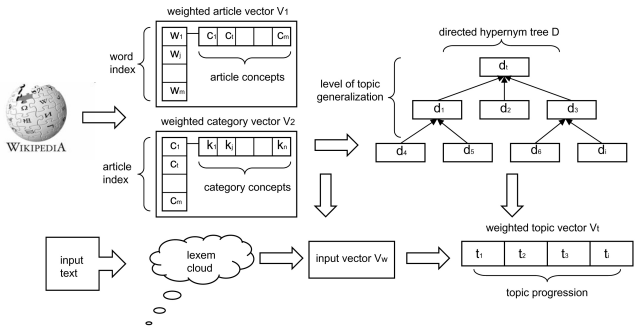

Figure 1. The system architecture of the topic generalization technique using the Wikipedia category taxonomy.

III. WIKIPEDIA SOCIAL SEMANTICS

Social networks such as Wikipedia offer the possibility of enhancing existing text representations through human-defined concepts. In this sense, concepts can be either reflected by Wikipedia articles or by their corresponding category information. The Wikipedia document collection is highly organized and constantly extended through the work of volunteers.

A. Wiki Feature Vector

Our approach utilizes both concepts in order to find the most prominent topics for a given document. An advantage of this combination is that even if a specific bit of information is not explicitly mentioned within an article set (for instance, when comparing two text fragments or words), we are able to predict their semantic relatedness on the basis of their shared category trails (implicit information). Our approach maps a minimum representation of the Wikipedia dataset into a weighted vector space using a rigorous feature reduction technique. That is, each article concept comprises an individual feature vector $V0$ with assigned token features w_i of strength k_i weighted by the TF-IDF scheme [28]. Like Gabrilovich and Markovitch (2007) [21], we additionally build an inverted vector index $V1$ with assigned affinity scores of article concepts to the corresponding features. Let

$v1_i$ be an vector of the inverted index of word w_i and its associated article concepts $c_j = \{c_t, ..., c_M\}$ as vector entries weighted by k_{jt}. M defines the total number of considered concepts. In order to reduce $V1$ to a minimum feature representation, we order all c_j of $v1_i$ by k_{jt} in descending order and remove those c_{jt} whose affinity score k_{jt} is less than five percent of the highest k_j. Therefore, for each input word w_i we can retrieve the corresponding entries of $v1_i$ utilizing $s1_i = M$ as the number of vector entries by c_j. Next, we build a second vector $V2$ connecting article and category concepts. Let $v2_i$ be a vector of article concept c_i with its associated category concepts $k_j = \{k_j, ..., k_N\}$, where N defines the total number of comprised category concepts. Therefore, $s2_i = N$ reflects the length of $v2_i$ - the number of comprised category concepts - for a given article concept c_i. Following this, we are able to retrieve the number of unique category concepts $s3_i$ for a given input word w_i by iterating over $v1_i$ and collecting k_j. In addition, given a set of words $W = \{w_i\}$ of length p we merge all $v1_i$ into V_w as an weighted vector, reflecting the association of article concepts ordered by their strength.

$$V_w = \sum_{i=m}^{p} v1_i \cdot k_i \tag{1}$$

Since this global feature vector is sorted in descending order, the first entries of our vector V_w correspond to those article concepts which fits best to a given input text. Therefore we are able to predict for a given text fragment of length p the best associated articles within the Wikipedia network based upon our vector representation. This will be evaluated in Section IV.

B. Wiki Topic Generalization

Since we are focused on the task of topic identification we have to define descriptive topic labels. In contrast to other approaches using keyword extraction and clustering techniques, we are utilizing the category taxonomy of the social network. Note that this taxonomy is defined as open-ended. Therefore, topic labels are changing over time[1], since the Wiki community constantly creates new article and category concepts. In principle, we connect a given text fragment to the most specific category concepts, proverbially taking an uphill walk within the taxonomy and 'dye' the trail we have visited. We have extracted the category taxonomy of Wikipedia in a top-down manner starting from the most generalized category (*Category:Contents*) and subsequently connected all subordinated categories to its superordinate nodes. Doing this, we forced the taxonomy in the representation of a directed tree D with one artificial root. The task of walking up the taxonomy therefore means walking along the hypernym edges of the category tree. Note that for each edge we pass, a topic generalization is comprised. Let V_t be a vector of generalized category topics t_i. Having V_w as our article concept vector which represents best the text fragment W we primarily iterate over all entries of V_w with length l and add all $v2_i$ (our category concepts assigned to $v1_i$) to V_t. Note that $v2_i$ inherits the feature weight k_i of $v1_i$.

$$V_t = \sum_{i=m}^{l} v2_i \cdot k_i \tag{2}$$

In a second step we perform the generalization by using the assigned $t_i \in V_t$ limited to f to query hypernym categories $d_t \in D$ and adding them to V_t. As the feature weight we use the value of k_i from our starting point $v1_i$. Already used t_i are 'dyed' in order to circumvent an overestimation of certain categories as well as to walk in a cycle. The f parameter allows us to adjust the extent of topic generalization. Since V_t is also sorted in descending order, with the most general topics occur at the beginning of our topic vector, the most specific at the end. See Table II for an example ranking of topic generalization. The question that arises in order to compute efficiently even larger documents is: which words or features of a document should be used for the category alignment process? All occurring wordforms? Those comprising each sentence or paragraph? A combination of both? Since the complexity in computing V_w is linear to the number of comprised tokens p of the input document W, we also need to apply a feature reduction technique to W. We propose a differentiated document representation on the basis of lexical chains [29] in order to reduce complexity but to retain the most valuable information features. We are following therefore the approach of Waltinger, Mehler, and Heyer (2008)[30] in extracting so called *lexeme clouds* as a representation of *topic chains*. Since lexical chains are derived by tracking and connecting semantically related tokens in a text, we need to be able to compute relatedness scores between individual token pairs.

C. Wiki Semantic Relatedness

We now present a measure for computing semantic relatedness – to derive a lexical chain representation – and its performance will be evaluated in section IV. Our approach utilizes concept frequency information of vector indices. This follows a search-engine-based word similarity distance, which is a measure derived from the number of returned hits for a given pair of keywords. The basic idea behind this approach is, that words with a similar or related meaning also tend to occur together in a given context. According to the work of Cilibrasi and Vitanyi (2007)[6] the Google distance (GD) can be derived by:

$$GD(x, y) = \frac{G(x, y) - \min(G(x), G(y))}{\max(G(x), G(y))} \tag{3}$$

[1]The method does not retrieve data directly from the original online repository (*http://download.wikimedia.org/*) but rather downloads snapshots of the data set within certain time slots.

Table II
TOP-5 SPECIALIZED AND GENERALIZED TOPIC CONCEPTS

Nowitzki's performance this year has vaulted him past the late Petrovic as the NBA's best-ever European import.
But the Bavarian Bomber is fast becoming one of the NBA's best players, period. Maybe even a little like Mike.
Last week, Dirk Nowitzki led the running, gunning Dallas Mavericks into the second round of the playoffs.
He put up, as the sports guys say, "Big-time numbers." Those would be: 100 points and 47 rebounds in a mere three games.

Related Articles	Specialized Topics	Generalized Topics
1. Dirk Nowitzki	1. basketball player	1. sport
2. Dallas Mavericks	2. basketball	2. United States
3. Avery Johnson	3. athlete	3. basketball
4. Jerry Stackhouse	4. olympic athlete	4. Germany
5. Antawn Jamison	5. basketball league	5. sport by country

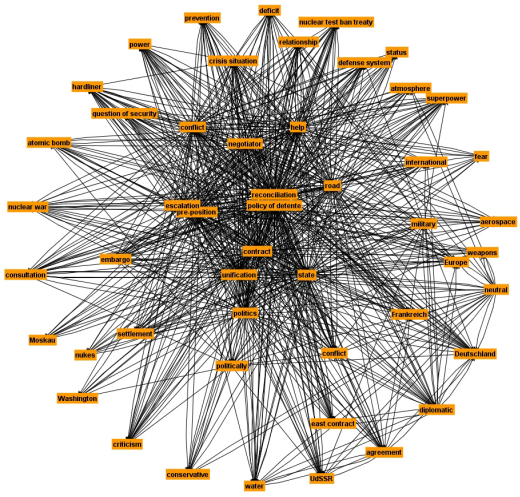

Figure 2. Connected graph representation of an input text. Nodes represent tokens of the text, edges represent the semantic relatedness scores obtained by WSR (Section III-C).

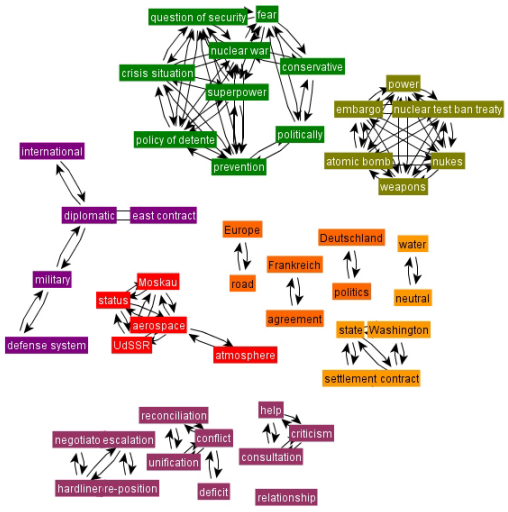

Figure 3. Lexical chaining representation of input text. The connected graph representation (see Figure 2) was decomposed by iterative graph clustering computing two iterations.

where the Google code of length $G(x)$ represents the shortest expected prefix-code word length of the associated Google event x. The expectation is taken over the Google distribution g. Adapting this idea to our reduced inverted index vector interpretation, we define the Wiki distance (WD) as:

$$WD(x,y) = \frac{\max\{\log(f_x), \log(f_y)\} - \log(f_{x,y})}{\log M - \min\{\log(f_x), \log(f_y)\}} \quad (4)$$

where f_x is the number of article concepts comprised by word x, f_y is the number of article concepts comprised by word y and $f_{x,y}$ is the number of unique articles comprised by x, y together. M is the total size of our index. Since we have two different vector representations (comprising article *and* category concepts) for a given word, we define $WD_{art}(x,y)$ as the Wiki distance obtained by the article index and $WD_{cat}(x,y)$ as the distance derived from category concept frequency information for the word pair x and y. Both distances are then combined, defining the Wiki Semantics Distance (WSD):

$$WSD(x,y) = \delta_i \cdot WD_{art}(x,y) + \delta_j \cdot WD_{cat}(x,y); \quad (5)$$

where δ is a weighting parameter. The combination is needed in order to measure even those word pairs that do not occur together in the article collection. Consider the following example: the word pair *colossus* and *gigant* do not occur together within the reduced article vector, hence they are related and connected through the category vector representation. In order to compute the semantic relatedness (WSR) of token pairs we invert the resulting distance.

$$WSR(x,y) = 1 - WSD(x,y) \quad (6)$$

where 0 defines no relatedness and 1 complete relatedness. See Table I for an example of relatedness scores for different domains.

D. Input Document Representation

In the last step we tackle the task of lexeme extraction of an input document or text fragment by computing lexical chaining. That is, we first convert the input document into a graph representation $G = (V, E, \sigma)$ where V is the set

of all used $w_i \in W$ and $E \subseteq V$ the corresponding set of edges and σ is the relatedness score computed by WSR (Figure 2). Next we decompose G into its main components following the incremental graph cluster algorithm proposed by [31]. In contrast to [31] we consider edge weights for the partitioning process, rather then the highest proportion of edges. As a result we gain a vector of sets of associated lexeme clouds L. Each $l_j \in L$ vary in the length of their entries reflecting the strength of topics in the document. See Figure 3 for an example lexeme cloud extracted from an article. For the category alignment process we select only the main component of L, which is the topic chain with the highest number of entries - the primary topic information.

IV. Empirical Evaluation

The calculation of our feature vector representation is based upon the German version of Wikipedia (February 2009). After parsing the XML dump comprising 756,444 articles we conducted the preprocessing by lemmatizing all input tokens and removing smaller concepts. We ignored those articles having fewer than five incoming and outgoing links and fewer than 100 non stopwords. The final vector representation comprised 248,106 articles and 620.502 lemmata. The category tree representation consisted of 55,707 category entries utilizing 128,131 directed hyponomy edges.

A. Corpora

In order to evaluate the performance of WSR a reference corpus of pre-classified word pairs is needed. For the German language there are – to our best knowledge – only two reference datasets with respective to the correlation to human judgment available. The first resource compiled by Gurevych (2005)[32] is a translation of the word-pair list initially created by Rubenstein & Goodenough [33] consisting of 65 word pairs. The second dataset was compiled by Cramer and Finthammer (2008) [34], comprising two lists of word pairs for which they obtained human judgements. The first list includes 100 word pairs - nouns manually collected from diverse semantic classes, e.g. abstract nouns, such as `knowledge` and concrete nouns, such as *flat-iron*. The second list comprises 500 word pairs which are part of collocations (e.g. *to help with words and deeds*) or association relations (e.g. *Africa and Tiger*). For the English language we used the *WordSimilarity-353* collection of Finkelstein, et al (2001) [35], since various results were reported. We make use of all these lists in order to cover a wide range of relatedness types and levels and to have a *true* comparison to current state-of-the-art approaches. As a second evaluation we used WSR in order to split compounds. In this scenario the relatedness score reflects the affinity between a number of connected nouns. We operate on a reference corpus manually created by Holz and Biemann (2008)[36], comprising 700 long German nouns. This dataset consists of 13 single nouns (no compounds), 640 two-part compounds, and 47 words consisting of three noun parts. In order to evaluate the topic identification on OTM we compiled two datasets comprising 1000 articles of the German Wikipedia (dataset B) and of the Meyer-Lexikon collection [2] (dataset A) each. Since both datasets are encyclopedia based and categorized by a taxonomy, we chose ten categories (e.g. fashion, politics, sports) as our open topics and selected for each category 100 articles. For each document, we computed the five and ten best generalized categories and compared if one of these matched the *initial*[3] category of the taxonomy. Note, we removed all category and HTML-markup information from the corpus. We report the level of accuracy this task was fulfilled. Additionally in the case of mismatches, we report on how close we are finally connected in the taxonomy. Doing this, we included a category context window of five levels. See Table IV-A for an overview of comprised sub-categories on the general topic level. Consider the following example: A article is finally categorized as a subordinate of our target category (level 1), when one of our five best topic labels matches one of the subordinate categories.

B. Results

Table VII
Results of 700 compound split experiment

Eval/Set	GS	CS	WSR
Precision	0.56	0.84	0.864
Recall	0.68	0.73	0.851
F1-Measure	.62	0.78	**0.857**

Overall our semantic relatedness approach performs very well. Tables IV and V show the results for all three German datasets comparing lexical network (e.g. Leachock & Chodorow, Hirst & St-Onge, Resnik), distributional (LSA, Google) and Wikipedia-based measures. Reference results were obtained by Cramer (2008)[17] and Gurevych (2005)[32]; results of the Latent Semantic Analyis (LSA) were kindly provided by Tonio Wandmacher [37]. The results of the Explicit Semantic Analysis (ESA1) were computed by applying the method of Gabrilovich and Markovitch (2007)[21] on the German and English Wikipedia document collection. Comparing the correlation results overall, we can see that the lexical network based measures show rather low coefficients. The distributional measures (LSA, Google) perform better. However, WSR outperforms all other relatedness scores except on the English dataset. The original ESA implementation reports a Spearmann correlation of 0.75 in contrast to WSR of 0.72. Our implementation of this method $ESA2$ shows a correlation of 0.70. This might be due to the selected

[2] http://lexikon.meyers.de/

[3] We considered only one category for each article even if Wikipedia articles are multiply categorized

Table III
NUMBER OF COMPRISED SUBCATEGORIES ON GENERAL TOPICS BY LEVEL

level/topic	info	spor	poli	medi	liter	cult	econ	mili	educ	cloth	relig
0	1	1	1	1	1	1	1	1	1	1	1
1	12	44	34	31	12	41	21	27	20	10	63
2	76	671	415	284	275	549	248	234	248	43	588
3	293	3339	1923	807	609	2760	1181	989	939	70	1781
4	631	8026	4403	1207	1075	6483	2995	1922	1951	81	2870
5	889	8618	8351	1483	1388	9930	4228	2440	2156	83	4083

Table IV
CORRELATIONS (*Pearson* COEFF. TO HUMAN ESTIMATES) TESTED FOR A AND B DATASET (GERMAN)

Test set	Leacock Chodorow	Jiang Lin	Hirst St-Onge	NSD *Google*	Semantic Vector	LSA (newspaper)	ESA1	WSR
r Set A	0.48	0.46	0.47	0.37	0.51	0.64	0.52	**0.77**
r Set B	0.17	0.25	0.32	0.36	0.28	0.63	0.44	**0.64**

Table V
CORRELATIONS (*Pearson* COEFF. TO HUMAN ESTIMATES) TESTED FOR GUR-65 DATASET (GERMAN)

Test set	Google	Lesk1 (DWDS)	Lesk2 (radial)	Lesk3 (hypernym)	Resn. Resnik	WND (Wiki)	ESA1	WSR
r Set $GUR-65$	0.59	0.53	0.55	0.60	0.72	0.71	0.56	**0.75**

Table VI
CORRELATIONS (*Spearman* COEFF. TO HUMAN ESTIMATES) TESTED ON WORD-SIMILARITY 353 DATASET (ENGLISH)

Test	WordNet	Roget's Thesaurus	WikiRelate	LSA	ESA-ODP	ESA1.	ESA2	WSR
r Set $WordSim-353$	0.35	0.55	0.48	0.56	0.65	0.70	**0.75**	0.72

snapshot of the Wikipedia dump or preprocessing-related differences. The results of the Semantic Vector are based on the implementation of Widdows and Ferraro (2008)[8] using the Wikipedia dataset, and gaining only mediocre results. Overall WSR performs best on all three German datasets gaining a Person correlation of up to .77. Comparing the results of Table VII we can observe that with an F-Measure of .857, WSR can also be successfully applied to the task of compound splitting. We retrieve better results on both recall and precision than those reported by Holz and Biemann (2008)[36]. The results of the last experiment are shown in Tables 8 – 11. As we can observe the Wiki Topic Generalization also performs very well with an average accuracy of .627 (level 0) and .705 (level 1) on the topic identification experiment. Since our approach utilized a Wikipedia dump the results on dataset B are not surprising. However, identifying the five/ten best topic labels out of a set of over 55,000 is still very good. Therefore, results on dataset A (1000 articles from Meyer-Lexikon) with an accuracy of .651 (level 0) and .731 (level 1) support the good performance on the topic identification task for OTM. Additionally, analyzing the results of the OTM we can observe there are many examples that are truly aligned on the right or the same article within Wikipedia, hence the generalization process over-generalized or under-generalized the category trails. Therefore most of the incorrectly-classified documents were actually labelled correctly – hence not exactly the one we defined and therefore marked as *false*. For instance, the article *cd burner* is tracked to the Wikipedia article concept *cd burner* but generalized to *technical instrument*, *storage medium*, *hardware* but not directly to *informatics* Moreover, since we randomly downloaded articles identified using the specific category information, we did not perform a corpus cleaning. That is, we did not remove inappropriate articles as disambiguation pages or miss-classified articles within the document collection. Overall our approach shows very promising results in generating specified and generalized topic labels for input documents.

V. CONCLUSIONS

We presented an approach using a social ontology to label documents by means of Open Topic Model. We considered the task of topic labelling as a task of document alignment within a social network. Comprised category information of article concepts was used to conduct a topic generalization. The performance was evaluated against two different corpora reporting an accuracy of up to .73. Additionally, we proposed a method for measuring semantic relatedness using a reduced vector representation of the Wikipedia document collection. Evaluation was performed using multiple human judgement experiment data sets gaining a correlation of up

Table VIII
ACCURACY ON OTM IDENTIFICATION MEYERS LEXIKON (10)

level/topic	info	spor	poli	medi	liter	cult	econ	pada	reli	psycho
A0	.638	.745	.750	.710	.660	.495	.710	.710	.760	.462
A1	.670	.798	.940	.770	.750	.546	.710	.810	.850	.527
A2	.766	.957	1	.860	.830	.825	.940	.920	.960	.714
A3	.798	.979	1	.860	.970	.979	.980	.940	.960	.725
A4	.872	.979	1	.890	1	.989	1	.950	.970	.725
A5	.894	.979	1	.910	1	1	1	.950	.970	.824

Table IX
ACCURACY ON OTM IDENTIFICATION MEYERS LEXIKON (5)

level/topic	info	spor	poli	medi	liter	cult	econ	pada	reli	psycho
A0	.553	.691	.680	.650	.640	.319	.580	.518	.710	.396
A1	.564	.702	.880	.720	.710	.392	.580	.639	.810	.439
A2	.691	.883	.980	.790	.800	.753	.880	.807	.920	.648
A3	.723	.936	.980	.800	.950	.969	.950	.928	.940	.659
A4	.777	.936	.980	.840	.990	.969	.980	.976	.950	.659
A5	.798	.936	.990	.880	.990	.979	.990	.976	.950	.747

Table X
ACCURACY ON OTM IDENTIFICATION WIKI LEXIKON (10)

level/topic	info	spor	poli	medi	liter	cult	econ	mili	educ	cloth
B0	.677	.630	.740	.660	.780	.520	.460	.620	.560	.240
B1	.768	.690	.880	.700	.850	.650	.490	.620	.710	.240
B2	.849	.870	.960	.810	.900	.970	.920	.890	.890	.280
B3	.879	.880	.970	.820	.960	.990	.990	.910	.950	.360
B4	.889	.900	.980	.830	1	1	.990	.930	.970	.360
B5	.929	.900	.990	.850	1	1	.990	.930	.980	.360

Table XI
ACCURACY ON OTM IDENTIFICATION WIKI LEXIKON (5)

level/topic	info	spor	poli	medi	liter	cult	econ	mili	educ	cloth
B0	.606	.590	.690	.520	.740	.350	.410	.520	.500	.140
B1	.717	.650	.810	.550	.790	.510	.420	.530	.640	.150
B2	.788	.730	.940	.650	.830	.900	.870	.820	.820	.170
B3	.808	.740	.950	.660	.910	.970	.930	.830	.930	.230
B4	.828	.750	.980	.680	.990	.980	.940	.870	.940	.230
B5	.889	.750	.990	.790	1	.980	.970	.870	.940	.230

to .77. For future work, we will focus on the task of topic generalization in depth. That is, conducting another feature weighting and reduction technique within the generalization process in order to assess the amount of generalization more precisely and to overcome an overestimation occurring in certain category trails.

ACKNOWLEDGMENT

We gratefully acknowledge financial support of the German Research Foundation (DFG) through the EC 277 *Cognitive Interaction Technology*, the SFB 673 *Alignment in Communication* (X1), the Research Group 437 *Text Technological Information Modeling*, the DFG-LIS-Project *P2P-Agents for Thematic Structuring and Search Optimization in Digital Libraries* and the *Linguisitc Networks* project funded by the German Federal Ministry of Education and Research (BMBF) at Bielefeld University.

REFERENCES

[1] D. Widdows, *Geometry and Meaning*. Center for the Study of Language and Inf, November 2004.

[2] E. Terra and C. L. A. Clarke, "Frequency estimates for statistical word similarity measures," pp. 244–251, 2003.

[3] F. Keller and M. Lapata, "Using the web to obtain frequencies for unseen bigrams," *Computational Linguistics*, pp. 459–484, 2003.

[4] M. Li, J. H. Badger, X. Chen, S. Kwong, P. Kearney, and H. Zhang, "An information-based sequence distance and its application to whole mitochondrial genome phylogeny." *Bioinformatics*, vol. 17, no. 2, pp. 149–154, 2001.

[5] R. Cilibrasi and P. Vitanyi, "Automatic meaning discovery using google," in *Manuscript, CWI, 2004; http://arxiv.org/abs/cs.CL/0412098*, 2004.

[6] R. Cilibrasi and P. M. B. Vitanyi, "The google similarity distance," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, p. 370, 2007.

[7] S. Deerwester, S. Dumais, T. Landauer, G. Furnas, and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society of Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[8] D. Widdows and K. Ferraro, "Semantic vectors," in *Proceedings of the LREC 2008*, E. L. R. A. (ELRA), Ed., Marrakech, Morocco, May 2008.

[9] C. Fellbaum, Ed., *WordNet. An Electronic Lexical Database*. The MIT Press, 1998.

[10] J. Ellman, "Eurowordnet: A multilingual database with lexical semantic networks: Edited by piek vossen." *Nat. Lang. Eng.*, vol. 9, no. 4, pp. 427–430, 2003.

[11] L. Lemnitzer and C. Kunze, "Germanet - representation, visualization, application," in *Proceedings of the 4th Language Resources and Evaluation Conference*, 2002, pp. 1485–1491.

[12] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," in *WordNet: An Electronic Lexical Database*, C. Fellbaum, Ed. The MIT Press, 1998, pp. 265–284.

[13] Z. Wu and M. Palmer, "Verb semantics and lexical selection," in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, 1994, pp. 133–138.

[14] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy," in *Proceedings of the IJCAI 1995*, 1995, pp. 448–453.

[15] D. Lin, "An information-theoretic definition of similarity," in *Proceedings of the 15th International Conference on Machine Learning*, 1998, pp. 296–304.

[16] G. Hirst and D. St-Onge, "Lexical chains as representation of context for the detection and correction malapropisms," in *WordNet: An Electronic Lexical Database*, C. Fellbaum, Ed. The MIT Press, 1998, pp. 305–332.

[17] I. Cramer, "How Well Do Semantic Relatedness Measures Perform? A Meta-Study," in *Semantics in Text Processing. STEP 2008 Conference Proceedings*. College Publications, 2008, pp. 59–70.

[18] D. Milne, "Computing semantic relatedness using wikipedia link structure," in *Proc. of NZCSRSC07*, 2007.

[19] S. Ponzetto and M. Strube, "Wikirelate! computing semantic relatedness using wikipedia," July 2006.

[20] T. Zesch, I. Gurevych, and M. Mhlhuser, "Comparing wikipedia and german wordnet by evaluating semantic relatedness on multiple datasets," in *In Proc. of NAACL-HLT*, 2007.

[21] E. Gabrilovich and S. Markovitch, "Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis," *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pp. 6–12, 2007.

[22] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," in *Proc. KDD Workshop on Text Mining*, 2000.

[23] H. Li and K. Yamanishi, "Topic analysis using a finite mixture model," *Inf. Process. Manage.*, vol. 39, no. 4, pp. 521–541, 2003.

[24] C. Wartena and R. Brussee, "Topic detection by clustering keywords," in *DEXA '08: Proceedings of the 2008 19th International Conference on Database and Expert Systems Application*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 54–58.

[25] Y. Yang, "An evaluation of statistical approaches to text categorization," *Journal of Information Retrieval*, vol. 1, pp. 67–88, 1999.

[26] N. Choi, I.-Y. Song, and H. Han, "A survey on ontology mapping," *SIGMOD Rec.*, vol. 35, no. 3, pp. 34–41, 2006.

[27] A. Mehler and A. Storrer, "What are ontologies good for?" in *Proceedings of OTT'06 — Ontologies in Text Technology*, U. Mönnich and K.-U. Kühnberger, Eds., Osnabrück, 2007, pp. 11–18.

[28] G. Salton and M. McGill, *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.

[29] A. Mehler, "Lexical chaining as a source of text chaining," in *Proc. of the 1st Computational Systemic Functional Grammar Conference, Sydney*, 2005.

[30] U. Waltinger, A. Mehler, and G. Heyer, "Towards automatic content tagging: Enhanced web services in digital libraries using lexical chaining," in *WEBIST 2008, 4-7 May, Funchal, Portugal*, Barcelona, 2008.

[31] D. Widdows and B. Dorow, "A graph model for unsupervised lexical acquisition," in *19th International Conference on Computational Linguistics, Taipei (COLING 2002)*, August 2002, pp. 1093–1099.

[32] I. Gurevych, "Using the structure of a conceptual network in computing semantic relatedness," in *Proceedings of the IJCNLP 2005*, 2005, pp. 767–778.

[33] H. Rubenstein and J. B. Goodenough, "Contextual correlates of synonymy," *Communications of the ACM*, vol. 8, no. 10, pp. 627–633, 1965.

[34] I. Cramer and M. Finthammer, "An evaluation procedure for word net based lexical chaining: Methods and issues," in *Proceedings of the 4th Global WordNet Meeting*, 2008, pp. 120–147.

[35] L. Finkelstein, E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin, "Placing search in context: the concept revisited." in *WWW*, 2001, pp. 406–414.

[36] F. Holz and C. Biemann, "Unsupervised and knowledge-free learning of compound splits and periphrases," in *Proceedings of CICLing 2008, LNCS 4919*, A. Gelbukh, Ed. Springer, 2008, pp. 117–127.

[37] T. Wandmacher, "How semantic is Latent Semantic Analysis?" in *Proceedings of TALN/RECITAL'05*, Dourdan, France, 2005.