ONTOLOGY SWITCHING FOR THE SOCIAL SCIENCES.
METHODS FOR THE UNDERLYING CORPUS ANALYSIS

Robert Strötgen[1]
University of Hildesheim

Semantic heterogeneity can be defined as the use of different documentation languages or ontologies for different document sets in digital libraries or other integrated document collections. Automatic translation between terms of different ontologies ("ontology switching") is one approach to solve this problem. The underlying statistical methods based on corpus analysis allow the automatic identification of semantic relations (or "cross-concordances") between terms from different ontologies.
These methods require parallel corpora with documents that are indexed with terms from different ontologies. For the social sciences suitable corpora are very rare. Therefore, special modifications of the ontology switching approach help to deal with this lack. One alteration is simulating a parallel corpus for un-indexed document collections using machine learning and linguistic methods. Another change is translating not between two ontologies but between an ontology and suitable free text terms.

Key words. Semantic Heterogeneity, Machine Learning, Natural Language Processing, Digital Libraries, Metadata

1 TREATMENT OF SEMANTIC HETEROGENEITY

Semantic heterogeneity in our context occurs e.g. when resources in a distributed digital library are indexed with different ontologies or subject schemes. It is even more lucidly if documents indexed with structured metadata from a bibliographic database meet completely un–indexed documents from the Internet.

One approach is expediting standardization efforts such as the *Dublin Core Metadata Initiative* (DCMI)[2], but they are only able to relieve the problem, not to solve it.
The project *CARMEN* as part of the German "Global Info"-Programme dealt with this kind of heterogeneity with different approaches: Support the use of metadata for scientific Internet documents, forward the standardization using the *Dublin Core Element Set* and the *Resource Desription Framework* (RDF)[3] and handle the remaining semantic heterogeneity with a set of intellectual and automatic methods. These were

- metadata extraction from un–indexed Internet documents,
- creating cross–concordances between different subject schemes intellectually,

[1] stroetgen@uni-hildesheim.de
[2] http://dublincore.org/
[3] http://www.w3.org/RDF/

- computing semantic relations between terms from different documentation
- languages by corpus analysis (cf. section 2 for more details), and
- query translation using these cross-concordances and semantic relations.

The results of a first simple evaluation were rather promising. It had shown that the query translation using statistically created semantic relations lead to new relevant documents for a set of test queries creating some sustainable noise. (Strötgen, 2002b, 2002c)
But there also were some open questions and issues:

- How improve the creation of semantic relations based on parallel corpus analysis?
- How solve some (linguistic) problems concerning the special case combining free text terms with (controlled) documentation languages?
- How transfer the methods evaluated within the domains Mathematics, Physics and Social Sciences to other domains?
- How allow users to parameterize the query translation and evaluate end user queries in addition to the previous retrieval quality tests? (Strötgen, 2002a)

With the intent to answer at least some of these questions the work is continued in the project ASEMOS[4]. First of all the problems concerning free text terms and parallel corpus analysis are in progress.[5] Additionally the previous results will be transferred to the domain of patent information.

The work is related to other projects dealing with semantic relations and (parallel) corpus analysis. A similar approach of using relations between descriptors have been used in the project *AIR/PHYS*. (Biebricher, Fuhr, Lustig, Schwantner, & Knorz, 1988) The *"EuroSpider"*[6] systems use related methods for multilingual databases. (Braschler & Schäuble, 2000; Schäuble, 1988) Such methods have also been investigated in the "Interspace"[7] prototype. (Chang & Schatz, 1999; Chung, He, Powell, & Schatz, 1999)

The work is also related to many projects in the context of the *Semantic Web*[8] and the use of ontologies.[9] Some projects deal with the learning of ontologies. (Maedche, 2002) Other projects try to integrate semantic heterogeneity on data level. (Bornhövd, 2001)

One of the first projects experimenting with automatic switching between controlled vocabularies was the rule based "Subject Switching" at the NASA in 1983. (Silvester & Klingbiel, 1993) This approach has been progressed in other projects like *ELVIRA* (Hellweg, 2002), *ViBSoz* (Marx & Müller, 2001) and *MyShelf*. (Hanke, Mandl, & Womser-Hacker, 2002; Kölle, Mandl, Schneider, & Strötgen, 2004)

[4] *A*pplied *S*tatistics, *E*valuation and *M*achine learning: *O*ntology *S*witching

[5] The work on the domain of Social Sciences is done in cooperation with the Social Science Information Centre, Bonn (http://www.gesis.org/iz/). (Binder et al., 2002; Hellweg et al., 2001; Krause, 2003)

[6] http://www.eurospider.ch/

[7] http://www.canis.uiuc.edu/projects/interspace/

[8] http://www.w3.org/2001/sw/

[9] e.g. the *WebOnt Working Group* in the *Semantic Web* context (http://www.w3.org/2001/sw/WebOnt/).

2 CORPUS ANALYSIS AND SEMANTIC RELATIONS

The computing of semantic relations in *CARMEN* is based on the conditional probability and the equivalence index. (Strötgen, 2002b) Therefore the tool *JESTER* is used that generates statistical correlated relations based on parallel corpora. (Hellweg, 2002) The result of the corpus analysis is a term-term-matrix, which is used for query translation later on.

The precondition to compute semantic relations is a parallel corpus, that is indexed with two documentation languages. In the context of *CARMEN*, which tried to link one documentation language with free text terms, these free text terms are treated as a documentation language.[10] The problem is that for the Social Sciences there is a lack of Internet documents indexed with a documentation language. Therefore a parallel corpus is simulated by assigning terms from a documentation language (the "Thesaurus Sozialwissenschaften"[11]). In *CARMEN* this assignment was done by using a probabilistic search engine and indexing the training collection of Internet documents from the Social Sciences. The ranking value for a thesaurus term supplied by the search engine was used as weight for the keyword assignments.

This method is obviously rather weak because keywords are only assigned to documents, which contain exactly this keyword. To improve the simulation of parallel corpora methods of machine learning and automatic classification are used in *ASEMOS*. For the experiments the data mining libraries of *WEKA*[12] are used. (Witten & Frank, 2001) Classifiers were trained using the *GIRT*[13] collection of bibliographic reference records of the Social Sciences. Full text terms were extracted from the title and abstract, and thesaurus terms were assigned intellectually.

The number of free text terms is enormous, what makes automatic classification with normal resources impossible. Therefore some linguistic preprocessing was necessary. The free text terms were stemmed using a stemmer for the German language from the *Apache Lucene* project.[14] But the reduction of the term count was not sufficient. A part of speech tagger[15] was used to select only some suitable word classes like nouns, adverbs, adjectives and full verbs. The POS tagger was trained with the *NEGRA* corpus,\footnote[16] a syntactically annotated corpus of German newspaper texts containing 20,602 sentences and 355,096 tokens, before. The newspaper genre seemed to be rather qualified for a Social Science corpus. Additionally named entities were extracted that were not processed by the stemmer.

From the about 13,000 documents from the *GIRT* collection only a small part could be used for training the classifiers because of resource limitations. Different classifiers from the *WEKA* suite were tested like instance based classifiers, support vector machine and Naive Bayes. Further different attribute types for the free text terms (nominal, term frequency and tfidf) were used. Particularly instance based classifiers were rather promising, but the resources allowed only the use of a very small fraction of the training corpus; support vector machine based classifiers had similar problems. Naive Bayes based classifiers hat a much better performance in training and classifying but lead to poorer results. But due to the projects resources the Bayes

[10] As a side effect free text terms requires some linguistic processing contrary to a controlled documentation language.

[11] http://www.gesis.org/en/information/support/

[12] http://www.cs.waikato.ac.nz/~ml/weka/

[13] http://www.gesis.org/Forschung/Informationstechnologie/CLEF-DELOS.htm

[14] http://jakarta.apache.org/lucene/

[15] http://web.bham.ac.uk/O.Mason/software/tagger/

[16] http://www.coli.uni-sb.de/sfb378/negra-corpus/

based classifiers were used for further experiments.

The trained classifiers were used to classify the about 7,000 Internet documents (only html) from the *CARMEN* corpus from the Social Sciences. The html documents were preprocessed by *JTidy*[17] to clean the code and create DOM documents. From the DOM documents the text of the title and all body text nodes were extracted with XPath queries using *Apache Xalan*[18].

Another weakness of the *CARMEN* implementation was that only terms from the documentation language were used as “free text terms”. This was a concession to the projects resources but had of course a very bad impact: Query translations from free text terms to thesaurus terms were very rare. As a consequence in *CARMEN* only the translation from the thesaurus to “free text terms” was tested. The linguistic preprocessing described above (POS tagging and stemming) allows the use of „real“ free text terms and the use of translations from free text to a thesaurus.

3 RESULTS

For evaluating the quality of the *WEKA* classifier results the classified documents were compared with the intellectually assigned classes for the training set. (cf. Fig. 1) The classifiers were trained with a random sample of 1,000 to 5,000 documents from the *GIRT* collection. Another random sample of 100 documents was used to evaluate the classifiers.

For each classifier the correct and false positives and negatives compared to the intellectual classification were used. This is a rather weak method because intellectual classification of documents is usually not very exhaustive and false positives might however fit to the documents. E.g. a document intellectually indexed with the subjects (among others) “China“ and “cross-cultural communication“ was classified automatically with the classes “Asia“ and “East Asia“ as well as “culture“. These all were counted as false positives in the automatic comparison but are classified pretty correctly. A spot check showed that e.g. for the Naive Bayes classifier with discretize option about half of all false positives are rather suitable to the document. This moderates the large number of “false positives“ in Fig. 1.

classifier	*type*	*# train.*	*subjects*	*true pos.*	*false pos.*	*false neg.*	*c/f pos.*
NaiveBayes	nom.	1,000	5.75	0.56	5.19	5.77	13.00%
NaiveBayes	nom.	5,000	0.48	0.04	0.44	8.25	1.21%
NaiveBayes	tf	1,000	0.64	0.46	0.18	5.71	25.00%
NaiveBayes	tfidf	1,000	266.05	4.00	262.05	2.03	1.62%
NaiveBayes	tfidf + discr.	1,000	5.28	1.80	3.48	4.93	68.03%
NaiveBayes	tfidf + kernel	1,000	79.44	1.99	77.45	4.28	6.59%
NaiveBayes	tfidf	5,000	0.05	0.06	1.45	8.17	4.44%

Figure 1. Average Results for Automatic Classifiers in *ASEMOS*

[17] http://jtidy.sourceforge.net/

[18] http://xml.apache.org/xalan-j/

Further all Internet documents from the *CARMEN* corpus were classified with some of the previous generated classifiers. The result was compared with the result of the *CARMEN* classification (cf. section 1). Both Naive Bayes classifiers with tfidf weighted attributes assigned less subject classes to the documents, but they differ rather strong: The classifier with kernel estimation has more assigned classes but a distinct smaller overlap with the *CARMEN* classifier. (cf. Fig. 2)

classifier	***type***	***av. subjects***	***SD***	***av. overlap with CARMEN***
CARMEN		39.75	71.29	./.
NaiveBayes	tfidf + discr.	7.91	13.84	3.05
NaiveBayes	tfidf + kernel	15.68	25.25	0.61

Figure 2. Compare *ASEMOS* Classifiers with *CARMEN* Classifier

The result of the following co--occurrence analysis with *JESTER* are term-term-matrices with semantic relations between subjects from the thesaurus and a selection of (stemmed) free text terms. These are very hard to compare with the *CARMEN* matrices that do not contain “real“ free text terms. Of course much more terms are related to a subject class in *ASEMOS* than in *CARMEN*.

The assignment of less subject classes to one document might lead to a slightly more precise computation for semantic relations. But the impact of the new classification and linguistic preprocessing needs much more evaluation yet.

4 OUTLOOK

It has shown that there is some impact on the corpus analysis using methods of machine learning and more sophisticated linguistic preprocessing. But the impact on the retrieval quality is not clear at all yet.

Some work on optimization is needed, but this takes much time, because every run needs some days computation time. Optimization of the machine learning parameters needs to be connected with the retrieval quality evaluation.

A new chance to improve quality is using the new *SozioNet* corpus[19] that contains several hundreds of Internet documents from the Social Sciences intellectually indexed with the “Thesaurus Sozialwissenschaften”.

Additionally the results will be transferred to the domain of patent documentation in cooperation with the FIZ Karlsruhe.[20]

[19] http://www.sozionet.org/

[20] http://www.fiz-karlsruhe.de/

REFERENCES

Biebricher, P., Fuhr, N., Lustig, G., Schwantner, M., & Knorz, G. (1988). The Automatic Indexing System AIR/PHYS. From Research to Application. In Y. Chiaramella (Ed.), *SIGIR'88, Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Grenoble, France, June 13-15, 1988* (pp. 333-342): ACM.

Binder, G., Marx, J., Mutschke, P., Strötgen, R., Plümer, J., & Kokkelink, S. (2002). *Heterogenitätsbehandlung bei textueller Information verschiedener Datentypen und Inhaltserschließungsverfahren*. Bonn: IZ Sozialwissenschaften.

Bornhövd, C. (2001). *Semantikbeschreibende Metadaten zur Integration heterogener Daten aus dem Internet*. Aachen: Shaker.

Braschler, M., & Schäuble, P. (2000). Using Corpus-Based Approaches in a System for Multilingual Information Retrieval. *Information Retrieval, 3*(3), 273-284.

Chang, C. T. K., & Schatz, B. R. (1999). Performance and implications of semantic indexing in a distributed environment. In *Proceedings of the eighth international conference on Information and knowledge management* (pp. 391-398): ACM Press.

Chung, Y.-M., He, Q., Powell, K., & Schatz, B. R. (1999). Semantic indexing for a complete subject discipline. In *Proceedings of the fourth ACM conference on Digital libraries* (pp. 39-48): ACM Press.

Hanke, P., Mandl, T., & Womser-Hacker, C. (2002). Ein "Virtuelles Bibliotheksregal" für die Informationswissenschaft als Anwendungsfall semantischer Heterogenität. In C. a. W. Womser-Hacker, Christian and Hammwöhner, Rainer (Ed.), *Information und Mobilität: Optimierung und Vermeidung von Mobilität durch Information; Proceedings des 8. Internationalen Symposiums für Informationswissenschaft (ISI 2002)* (pp. 289-302). Konstanz.

Hellweg, H. (2002). Einsatz von statistisch erstellten Transferbeziehungen zur Anfrage-Transformation in ELVIRA. In J. Krause & M. Stempfhuber (Eds.), *Integriertes Retrieval in heterogenen Daten. Text-Fakten-Integration am Beispiel des Verbandinformationssystems ELVIRA* (Vol. 4). Bonn: IZ Sozialwissenschaften.

Hellweg, H., Krause, J., Mandl, T., Marx, J., Müller, M. N. O., Mutschke, P., et al. (2001). *Treatment of Semantic Heterogeneity in Information Retrieval*. Bonn: IZ Sozialwissenschaften.

Kölle, R., Mandl, T., Schneider, R., & Strötgen, R. (2004). Weiterentwicklung des virtuellen Bibliotheksregals MyShelf mit Semantic Web-Technologie: Erste Erfahrungen mit informationswissenschaftlichen Inhalten. In M. Ockenfeld (Ed.), *Information Professional 2011. Strategien - Allianzen - Netzwerke; 26. Online-Tagung der DGI, Frankfurt am Main, 15. bis 17. Juni 2004; Proceedings* (pp. 111-124). Frankfurt am Main.

Krause, J. (2003). *Standardisierung von der Heterogenität her denken. Zum Entwicklungsstand bilateraler Transferkomponenten für digitale Fachbibliotheken* (Vol. 28). Bonn: IZ Sozialwissenschaften.

Maedche, A. (2002). *Ontology Learning for the Semantic Web*. Boston et al: Kluwer Academic Publishers.

Marx, J., & Müller, M. N. O. (2001). The Social Science Virtual Library Project. Dealing with Semantic Heterogeneity at the Query Processing Level. In *Third DELOS Network of*

Excellence Workshop "Interoperability and Mediation in Heterogeneous Digital Libraries". Darmstadt, Germany, September 8-9, 2001 (pp. 19-23). Darmstadt.

Schäuble, P. (1988). An Information Structure dealing with Term Dependence and Polysemy. In Y. Chiaramella (Ed.), *SIGIR'88, Proceedings of the 11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Grenoble, France, June 13-15, 1988* (pp. 519-533): ACM.

Silvester, J. P., & Klingbiel, P. H. (1993). An operational system for subject switching between controlled vocabularies. *Information Processing and Management: an International Journal, 29*(1), 47-59.

Strötgen, R. (2002a). Behandlung semantischer Heterogenität durch Metadatenextraktion und Anfragetransfer. In C. Womser-Hacker, C. Wolff & R. Hammwöhner (Eds.), *Information und Mobilität: Optimierung und Vermeidung von Mobilität durch Information; Proceedings des 8. Internationalen Symposiums für Informationswissenschaft (ISI 2002)* (pp. 259-271). Konstanz: UVK.

Strötgen, R. (2002b). Meta-Data Extraction and Query Translation: Treatment of Semantic Heterogeneity. In M. Agosti & C. Thanos (Eds.), *Research and Advanced Technology for Digital Libraries: 6th European Conference, ECDL 2002, Rome, Italy, September 16-18, 2002; Proceedings* (pp. 362-373). Berlin: Springer.

Strötgen, R. (2002c). Treatment of Semantic Heterogeneity using Meta-Data Extraction and Query Translation. In W. Adamczak & A. Nase (Eds.), *Current Research Information as Part of Digital Libraries and the Heterogeneity Problem: Integrated Searches in the Context of Databases with different Content Analyses* (pp. 41 - 49). Kassel: Kassel University Press.

Witten, I. H., & Frank, E. (2001). *Data Mining. Praktische Werkzeuge und Techniken für das maschinelle Lernen*. München: Hanser.