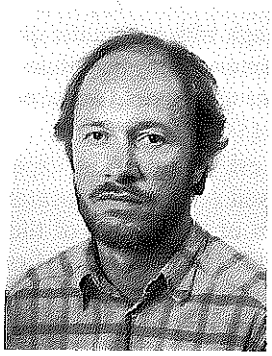
Gerda Ruge
Christoph Schwarz
SNI AP 312, Siemens AG, Munich, FR Germany



TERM ASSOCIATIONS AND COMPUTATIONAL LINGUISTICS

Ruge, G., Schwarz, Ch.: **Term associations and computational linguistics.**
Int.Classif.18(1991)No.1, p.19 - 25, 20 refs.
Most systems for term associations are statistically based. In general they exploit term co-occurrences. A critical overview about statistical approaches in this field is given. A new approach on the basis of a linguistic analysis for large amounts of textual data is outlined. (Authors)

1. Introduction

Searching textual documents has always faced some fundamental problems. They are caused by the rigid architectures and the limited possibilities of free-text search systems. Document indexing in information retrieval means,roughly speaking, either to use a controlled or uncontrolled vocabulary for intellectual indexing or to take nearly every word from the database as a descriptive entity in free-text indexing.

Free-text indexing by means of inserting everyword from the text into an inverted file is often a crude and superficial approach in view of the goal of catching the content of documents. Inverted files only offer limited search facilities by means of Boolean operators, distance measures between search terms and truncation of search terms.

The two main problems in searching textual douments by selecting search terms remain. First, relevant documents that contain different terms than the selected ones are not retrieved. Second, documents that contain the search term in an unwanted context are retrieved. Salton (1) pointed out that term association (TA) techniques may offer progress in this field though he does not yet see large scale working systems for TA.

2. Term association techniques

Previous approaches for term association are discussed in what follows. The theoretical foundations and first implementations go back some 20 years. New approaches that have been tried in the last years are then outlined.

2.1 Statistical approaches

In the sixties and seventies a lot of experiments in the field of TA were done. According to Giuliano (2) there are *first order TA* and *second order TA*. First order TA means that one exploits the fact that terms often occur in the same contexts. In this case they are thought to be semantically compatible, "they have to do with one another". They convey new ideas with regard to the use of other terms. Second order TA according to Giuliano (2) concentrates on terms that are interchangeable with regard to sets of contexts.

For the extraction of these TA no linguistic techniques are used, i.e. the internal structure of the contexts is not considered. In first order TA the number of contexts, in which two terms cooccur and the number of contexts, in which just one term occurs, is used as a basis for the calculation of their compatibility. This calculation is characterized by different similarity measures which yield a value between 0 and 1, the best value being 1.

The results of first order TA are being used as a basis for second order TA. It is assumed that terms are semantically similar if they are compatible with the same set of terms, i.e. if they have similar contextual environments. Second order TA therefore yields values that can be interpreted as measures for "semantic similarity" and not for "semantic compatibility" as in first order TA.

Most of the TA systems are first order TA systems. In order to build co-occurrency statistics very different entities are chosen as contexts and terms. In general the context of a document is considered. Sometimes this is a small set of manually built index terms (Sparck-Jones (3), p.72) or all terms (Rieger (4)) of the document. Moskovich (5) gives a summary of activities in this field in the Soviet Union where often smaller contextual entities are used. The entities for which TA are built are e.g. inflected words (Giuliano (2)) or stems (Sparck-Jones (3), p.135). The quantity of documents analyzed in order to get TA is usually several hundreds of documents and several hundreds of terms (see e.g. Lesk (6) or Sparck-Jones (3)).

TA are intended to be used for automatic query expansion. The user may also regard them as proposals for further search terms in an interactive dialogue. There is a general agreement among different schools in information retrieval (see e.g. Salton (1) and Robertson (7)) that TA are important but they cannot be used for automatic search term expansion to a query. Robertson (7) reports on discouraging results in this field and Sparck-Jones (3), who did a lot of performance tests with

TA, says that their application area is a very limited one. Although with automatic query expansion the retrieval performance may be enhanced, it is especially the common terms that yield worse performance results (Sparck-Jones (3), p.196).

Lesk (6) made TA experiments where he not only observed retrieval performance but also made an evaluation of the results of the two TA approaches. He concluded that among the first order TA only 20 percent of all terms were related in a semantically significant way. Among the second order TA he did not find any synonym.

In our opinion the main problem of the statistical approach lies in the size of the context chosen. Giuliano (2) points out that the contextual environment should be chosen very carefully. It is a fact that in a larger context, e.g. within document limits, there may well occur terms that have no semantic relation whatsoever. In large contexts nearly any term may cooccur with any other in the same context. For this reason similarities between nearly any possible term pairs have to be calculated. This means that besides the poor quality of statistic TA this approach also needs a lot of processing time and evaluations in this area can only be performed for small quantities of terms and documents.

2.2 Recent approaches

Besides the purely statistical systems on the basis of term cooccurrence in the original data there are several approaches that try to extract semantic relations from monolingual dictionaries. People working in this area start from the basic idea that the words of lexical definitions can be interpreted as semantic features. The similarity of terms is then being determined by a measure that shows their degree of overlap. Shaikevich (8) uses any word that is not a function word as a semantic feature. Chodrow (9) only takes so called genus terms, which are in general the first words in a lexical definition. Das-Gupta (10) uses the Chodrow (9) method for the disambiguation of logical AND or OR in queries, and argues that it is successful for this purpose.

The production of TA by means of lexical definitions looks promising. Its disadvantage is that only very little information on terms is available An extreme case is exemplified by Chodrow (9) where just one genus term is used. Shaikevich (8) points out that small changes in the algorithm have very severe consequences for the robustness of the systems.

During the last years modern information technology is concerned with the problem of finding alternative search terms by using approaches from expert systems and connectionism. Güntzer (11) developed an expert system that by analyzing queries deduces which TA might be relevant. Wettler (12) tries to apply cooccurrence statistics not by means of similarity measures but by using connectionist models. Chodrow (9) have used connectionist models on the basis of manually built term relations. One of the main advantages of these models is that they are expected to be to a high degree error tolerant (see also Fahlman (13)).

3. Intelligent text processing in TINA

The development of natural language processing systems for free-text has been pursued for a long time at Siemens in Munich. In 1982 the TINA (Text-Inhalts-Analyse: text content analysis) project was started there, based on more than 5 years of previous research and development work in information and documentation.

The aim of TINA is to develop natural language processing software for information retrieval: free-text search and indexing, automatic abstracting, thesaurus building and natural language query handling (for an information retrieval system see Schwarz (14), for other applications in the field of documentation see Schwarz (15)). Existing sub-projects were tested on commercial free-text databases on mainframes and new sub-projects are being developed for main frame applications as well as for text handling in an office environment on a personal computer.

In its first stage TINA was concerned primarily with morphology, i.e. term normalization (lemmatization and stemming) for indexing purposes. During the last years TINA research and development work has focused on syntactic analysis for free-text documents and large textual databases. The actual TINA work is going to be oriented to the field of semantics and combines linguistic techniques with statistical ones.

3.1 A computational linguistic approach

In the previous chapters we outlined
- that the use of cooccurrence statistics causes great problems because of the large contexts that are used,
- that sets of semantic features look promising for the extraction of TA, but the sets that can be used are neither large enough nor are the systems sufficiently robust.

In what follows we propose a new approach on the basis of linguistic relations that are generated on the basis of free-text documents. The relations are syntactic ones. Their generation is described and their use is compared to earlier systems.

We use two kinds of syntactic relations: One is the head/modifier relation in noun phrases, the other is a relation between heads of noun phrases that are combinded by conjunctions, called conjunction relation. They will be explained in the following paragraphs.

3.2 Syntactic relation: head/modifier

Figure 1 shows the most common heads and modifiers for *storage*. The modifier *water in* figure 1, e.g. is extracted from various noun phrases like e.g. *water storage* or *storage and transport of fresh water*. A detailed explanation of the noun phrase extraction is given in

Schwarz (14), a general outline is given in the following chapter.

head/modifier relations
storage

modifiers	frequency	heads	frequency
data	258	tank	512
heat	215	device	499
energy	202	container	252
charge	117	position	239
water	106	chamber	224
information	104	unit	206
main	86	battery	189
memory	60	system	184
liquid	58	compartment	173
image	54	area	160
material	39	capacitor	158
signal	39	cell	146

Fig.1: Head/modifier relation for *storage* and their frequency in 200,000 PTO abstracts from the years 1983, 1984,1985

3.3 Syntactic relation: conjnnction

The conjunction relation is a relation between terms taken from noun phrases that are conjoined by *and* or *or*. In *dangers of milk transport in metal containers or glass bottles* the two heads of the respective noun phrases conjoined by *or*, i.e. *containers* and *bottles*, are in the conjunction relation. Figure 2 shows the conjunction relations between *container* and the most common terms for this relation with regard to *container* (for a detailed discussion see Ruge (16)).

conjunction relations
container

term	frequency
container	68
closure	46
cover	31
bottle	22
method	19
lid	18
member	16
material	16
portion	15
tank	15
article	15
vessel	14
pump	14
package	14
cap	13
tube	12
cup	11
device	11
box	11

Fig.2: Conjunction relations for container and their frequencies in 200,000 PTO abstracts

3.4 Hypotheses for syntax based term association (TA)

On the basis of these two relations, i.e. the head/modifier relation and the conjunction relation, we verified several hypotheses that we thought useful for our aims. One hypothesis says that the higher the overlap of the heads and modifiers is the higher is the probability that these terms belong to a set of semantically related terms (synonyms, antonyms, hyponyms, hyperonyms etc.), i.e. a common *semantic field*. The other hypothesis says that the number of identical pairs of words in the conjunction relation conveys a degree for the probability of their semantic similarity.

Both relations are based on syntactic dependency relations as outlined in chapter 4. The evaluation of the hypotheses mentioned above will be discussed in chapter 5. In chapter 6 we report on our first implementation of the system.

4. Syntactic analysis in TINA

The two basic relations that we use as a basis for our term associations within TINA are based on a syntactic analysis for large amounts of textual data. In view of the enormous masses of text and the variety of text quality, TINA uses a shallow syntactic analysis. The basic principles have been known for a long time in information science, but have seldom been systematically applied to large text files (see Sparck-Jones (17)). The linguistic analysis produces a representation of the structural relationships among words in a text.

4.1 Syntactic dependency relations

The linguistic relation which is extracted and represented is the dependency relation. The dependency relation indicates which words modify other words in phrases (for example, "fish" modifies "scaler" in figure 3a). The aggregate of all meaningful words in a phrase together with their dependencies is called the dependency graph of a phrase. Figure 4 shows the dependency trees of the phrases in figure 3.

The modified word at the root of the dependency tree is called the *head* of the phrase that is represented by this dependency structure. "scaler" is the head of the example in figure 3a, while "located" is the head of the sub-phrase "located on the drive shaft" in figure 3b. All words which are at the beginning of a dependency arc are called *modifiers*.

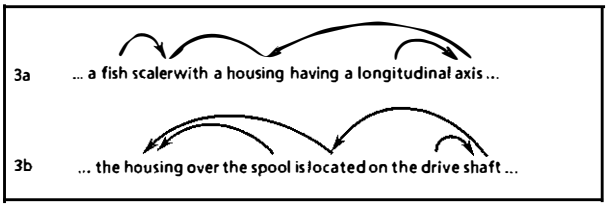


Fig.3: Specification relations in phrases

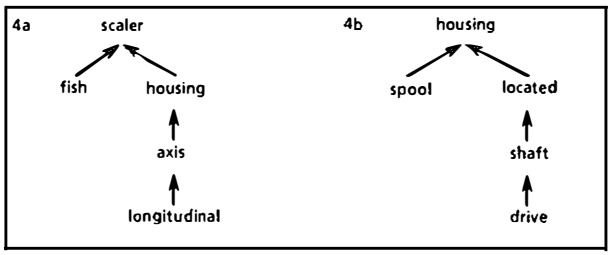


Fig.4: Dependency representation of the phrases in Figure 3

TINA's syntactic analysis produces dependency representations of all phrases in free texts. The analysis is designed with special emphasis on efficiency and effectiveness. Whether a rule is effective or not is evaluated empirically. New rules are developed at the points of greatest error rates and are implemented as efficiently as possible.

The first problem of generating dependency representations is the isolation of phrases in texts. This problem ranges from the examination of special characters (such as the punctuation mark .) to the recognition of phrase delimiters, such as subordinate conjunctions (e.g., *while*). The most difficult class of phrase delimiters, verbs, can only be recognized by context sensitive rules, because many words can be verbs in some context but adjectives or nouns in others.

The rules that determine the modifications are formulated in terms of syntactic categories, for example, noun, adverb, preposition etc. Before these rules are applied, the categories of the words have to be recognized.

The syntactic dependency analysis takes as its basis a well-known structural regularity in noun phrases (NP). This principle claims that parts of NPs that are related by prepositions are connected by a modification from right to left, while NPs built of parts related without prepositions are connected by a modification from left to right. Thus, "storage of milk" has its head ("storage") on the left, while "milk storage" has its head ("storage") on the right. However, there are many counter-examples to this basic principle, including some involving gerunds, adjectives (which may be postponed in certain contexts), and adverbs whose scopes are often difficult to recognize. All these irregularities require special context dependent syntactic rules that modify the above mentioned basic principle (for more details see Schwarz (14)).

During processing the syntactic relations are produced by invoking context patterns which are associated with the most significant categories or words that activate them. The text is processed from left to right and all rules which are activated by word categories are candidates for application. They are checked by taking the most probable first. The treatment of gerund construktions, for example, needs 10 context dependent rules. The context 'gerund followed by determiner', for example, invokes a syntactic relation to the gerund from the head of the sub-phrase that begins with the determiner and it invokes a modification from the head of the noun phrase that precedes the gerund to the gerund (see figure 5).

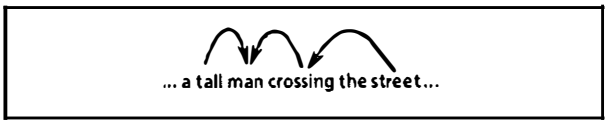


Fig.5: Modification relations in a noun phrase with a gerund

4.2 Some technical data of TINA's syntactic analysis

The TINA syntactic analysis for the generation of the dependency trees processes about 19 Kbyte of text in one CPU second on a Siemens BS2000 mainframe computer (4 MIPS). The processing of 200,000 abstracts from the US Patent and Trademark Office PTO (130 Mbyte of text) resulted in 140 Mbyte of dependency structures. The processing was done in 13 hours real-time on a BS2000 mainframe computer.

In order to evaluate the correctness of the syntactic analysis, we transferred the well-known information retrieval measures of recall and precision to the syntactic analysis. In particular, the "link recall", r, gives the rate of automatically extracted links within the document noun phrases. It corresponds to recall in information retrieval which gives a measure on the basis of the quantity of the retrieved documents. The "link precision", p, gives the rate of correct links automatically extracted. This is an analogy to information retrieval precision which gives a measure of the quality of the retrieved documents.

The evaluation was done by comparing intellectually man-retrieved links and links produced by the system. Let M be the "man-retrieved links" and S be the "system-produced links". Then recall and precision are defined in the usual way (see Salton (18), p. 164) by taking M as the set of relevant links and S as the set of retrieved links

$$r(M,S) = |M \cap S| \div |M|$$
$$p(M,S) = |M \cap S| \div |S|$$

We calculated a recall of 0.85 and a precision of 0.84 for the TINA syntactic analysis.

This error rate is due to the analysis of link *tokens*, which means one occurrence of one link in one text, e.g. the link between *data* and *storage* in some sentence of a document. That the link *type data storage* exists means that this link is possible. In the example of figure 1 258 tokens have the type *data storage*. With respect to the link types the error rate is not so high. And this again means that the above error rates of the link token generation in the syntactic analysis have hardly any influence on the hierachical odering of the link types in the head/modifier or conjunction relation tables (see figure1, figure2).

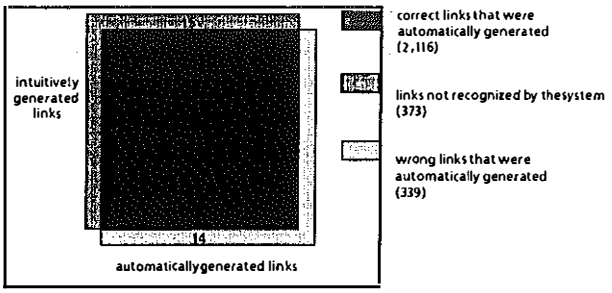


Fig.6: Correlation between intuitively generated links and automatically generated links

5. Syntactic relations as a basis for TA

Our basic idea was to consider TINA's syntactic relations as the context for cooccurrence statistics (see Ruge (19)). They can, in fact, be regarded as the smallest possible contexts, for the context consists of a single term only. In addition, by contrast to simple cooccurrence within document limits, there exists a semantic relation between terms that are gained by a linguistically motivated head/modifier or conjunction analysis.

In chapter 4 we outlined very roughly how syntactic dependency relations can be discovered in natural language texts. The totality of all these pair tokens gained either on the basis of the head/modifier or on the basis of the conjunction relation yields types of word pairs as shown in figure 1 and figure 2.

The head/modifier relations are already first order TA in the sense of Giuliano (2), for they exist only for terms that "have something to do with one another". They can serve as a basis for second order TA that among others contain synonymous relations. Terms that are interchangeable in these contexts, i.e. that have many heads and modifiers in common, are semantically similar. Our results (see below) suggest that the terms with many common heads and modifiers represent semantic fields if an adequate similarity measure can be defined for the sets of heads and modifiers. In addition, terms that are combined by the conjunction relation tend to be members of semantic fields.

The main problem of purely statistical TA approaches is the size of the contexts. The quantitiy is signifficantly reduced by using linguistically motivated contexts like the dependeny relations from TINA. Syntactic relations define very small contexts that guarantee semantic compatibility. But there is an additional advantage with respect to purely statistical properties of terms. There nearly any term can coocccur with any other within document limits. Therefore for an extremely high quantity of pairs the similarity must be calculated. By contrast, in the TINA approach this is not possible for all terms cooccur in syntactic relations. Otherwise grammatical selection restrictions would be violated. This means that a far larger amount of terms and documents can be handled with the same effort of computational cost.

5.1 Working hypotheses

As mentioned in chapter 3 we based our first implementation on two hypotheses. The first hypothesis says:
- The higher the correlation between the heads and modifiers of two terms, the higher the probability that the two terms belong to a common semantic field.

The second hypothesis says:
- The higher the frequency of occurrence of two terms in the conjunction relation, the higher the probability that the two terms belong to a common semantic field.

5.2 Verification of the first working hypothesis

For the verification of the first working hypothesis we calculated the correlation of every term in our 130 Mbyte patent database with every other term. All terms that occured in at least ten head or modifier types were taken into account for this purpose. The correlation that was calculated is the so-called S-coefficient.

$$(1)\ X \cap Y / X \cup Y$$

where X and Y are sets of the head or modifier types of the terms. This correlation was computed for the head and the modifer overlap of the terms seperately.

The set of possible correlations from 0 to 1 was partitioned into intervals. The interval [0.075,1] for example is associated with all pairs of terms that have a correlation of at least 0.075 and lower than 1 for head or modifier correlation respectively. These correlation intervals were defined in such a way that each interval contained twice the number of term pairs of the next highest interval. Thus the density of the term pairs in the intervals is decreasing exponentially. The following intervals of correlation were used for the examination of the first hypothesis: A = [0.075 , 0.1[, B = [0.1 , 0.125[, ..., K = [0.325 , 0.35[, L = [0.35 , 1].

From each of these intervals a set of about 100 term pairs was chosen arbitarily. These sets were judged intellectually according to wether their terms are semantically similar to a high degree, to a lower degree or not. Synonyms, antonyms or hyperomys were classified as high degree similarities. Term pairs that are semantically similar to a lower degree, e.g. pseudo-synonyms, were classified as having a probability of 0.5 of similarity.

The graph of figure 7 shows the probability of semantic similarity of term pairs dependent on the intervals of correlation for head correlation and for modifier correlation. According to this examination the probability of semantic similarity is increasing with increasing head or modifier correlation

5.3 Verification of the second working hypothesis

Term pairs that occur in the test database in conjunction relation were also examined. In a first step a random set of term pairs in conjunction relation were judged with respect to their semantic similarity. The conjunction pair types are already semantically similar with a probability of 0.4.

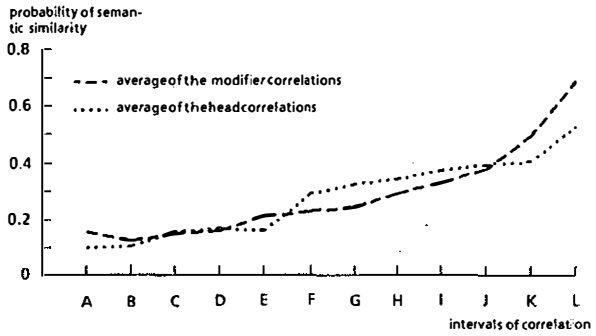


Fig.7: Probability of a common semantical field of two terms depending on their modifier and head correlation (average)

In a second examination the conjunction pair types were discriminated by their token frequencies in the test database. Arbitarily chosen sets of conjunction pairs, each having the same frequency, were again judged intellectually. Figure 8 shows the probability of semantic similarity dependent on their frequencies. The result of this examination was that the probability of semantic similarity is increasing with increasing frequency of the conjunction pairs. Thus both of our working hypotheses could be verified.

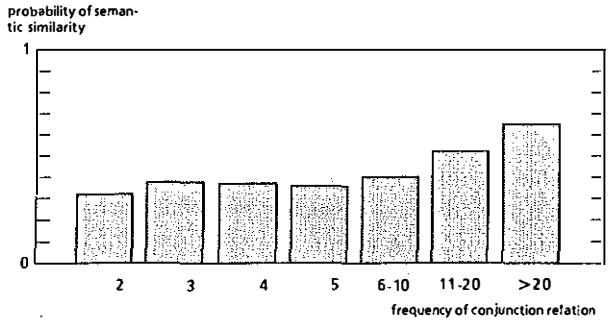


Fig.8: Probability of semantic similarity for term pairs in the conjunction relation

6. Implementation of a syntactically based TA system

Up until now we are not yet definitely sure which similarity measure we are going to use. At present we are trying several weighting strategies in order to improve our results. For the first implementation we decided in favour of a very plausible similarity measure.

As we expected that heads and modifiers represent some sort of properties of terms we chose a similarity measure that was based on the degree of overlap of contexts. It is important that the part of common properties is relatively high with regard to the total amount of properties (see Panyr (20), p. 55):

$$(2) \quad S(t_i,t_j) = (X_i \cap X_j) \div (X_i \cup X_j),$$

where Xi and Xj are the property sets, i.e. the heads and modifiers, for the terms ti and tj.

As (2) compares simple sets the frequency of tokens for a relation is not taken into consideration. We argue that for the determination of the semantic similarity it is important which properties a term may in principle have and not how frequent it is.

The sets of heads and modifiers had to be considered separately. As we could find no arguments that proved that heads were more important contexts than modifiers were, we weighted them equally. The similarity measure in (3) does not prefer either.

$$(3) \quad S(t_i,t_j) = (H_i \cap H_j + M_i \cap M_j)/(H_i \cup H_j + M_i \cup M_j)$$

where Hi and Hj are the sets of heads and Mi and Mj are the sets of modifiers for the terms ti and tj.

Some of the results of our first implementation are shown in figure 9. They were the result of processing 200,000 patent abstracts. As we knew from the results of our verification of the second working hypothesis, terms in the conjunction relation are good candidates for semantically related terms. Our first implementation was limited to a head/modifier correlation analysis built on the basis of these term pairs. An additional limitation of our first implementation was that we only considered the 30 most frequent heads and modifiers for a given term.

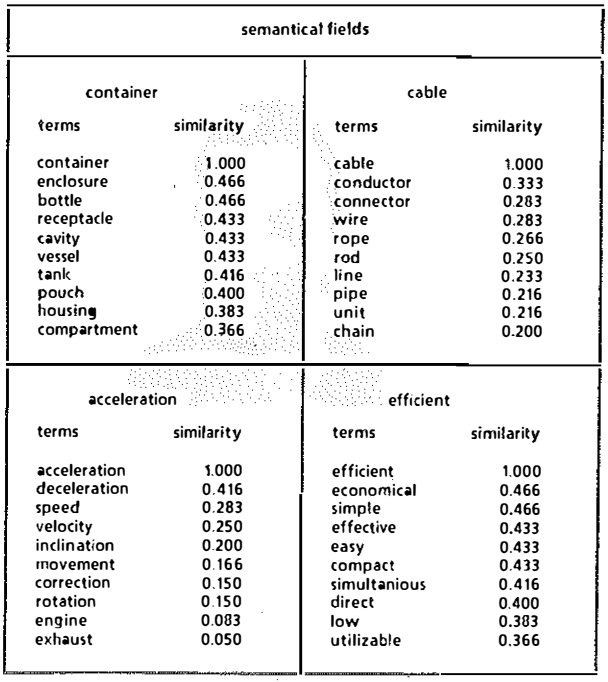
For 24,253 lemmatized terms there existed a conjunction relation. The word pairs were taken as a basis for the generation of the head/modifier relations. For a term there are at most 1,159 conjunction types and on the average 12 types. 66,049 terms appear in head/modifier relations. A single term can have up to 3,000 different head or modifier types. The semantically similar terms that are gained by the similarity measure (3) are arranged in figure 9 in descending order.

The semantic fields from our first implementation in figure 9 correspond more or less to what we expected, yet they did not seem satisfactory enough. Many important terms are still missing. This is mainly the case because our first implementation considered only terms that were already related by means of the conjunction relation. Therefore there appears no term that only differs from one another because of its writing variants. They will hardly ever be joined together by *and* or *or*. On the other hand there are terms that do not belong to a semantic field like *unit* and *cable*. Nevertheless, there are a lot of terms that really do belong to semantic fields: There are synonyms (e.g. *efficient* and *economical*), antonyms (e.g. *acceleration* and *deceleration*), hyperonyms (e.g. *container* and *bottle*), hyponyms (e.g. *cable* and *line*) and pseudo-synonyms (e.g. *cable* and *wire*).

7. On-going and future work

As yet we have no final solution to the integration of the similarity measure (2) and the conjunction relation. The verification of the second working hypothesis shows that among the term pairs that are related by the conjunction relation there is a high degree of semantic similarity (see figure 8). The integration of different weights for the head/modifier relation and the conjunction relation is expected to further improve the TA results.

In order to improve the quality of our semantic fields we are actually trying different similarity measu-



semantical fields

container

terms	similarity
container	1.000
enclosure	0.466
bottle	0.466
receptacle	0.433
cavity	0.433
vessel	0.433
tank	0.416
pouch	0.400
housing	0.383
compartment	0.366

cable

terms	similarity
cable	1.000
conductor	0.333
connector	0.283
wire	0.283
rope	0.266
rod	0.250
line	0.233
pipe	0.216
unit	0.216
chain	0.200

acceleration

terms	similarity
acceleration	1.000
deceleration	0.416
speed	0.283
velocity	0.250
inclination	0.200
movement	0.166
correction	0.150
rotation	0.150
engine	0.083
exhaust	0.050

efficient

terms	similarity
efficient	1.000
economical	0.466
simple	0.466
effective	0.433
easy	0.433
compact	0.433
simultanious	0.416
direct	0.400
low	0.383
utilizable	0.366

Fig.9: The first 10 entries of automatically generated semantical fields with similarity values

res. Several models are being tested. On the one hand the overlap of syntactic contexts is used as a model as outlined above. On the other hand we use a model that analyzes the position of terms in a semantic space that is defined by the contexts. In addition the influence of the amount of tokens in a syntactic relation is analyzed and the importance of types of relations. Tokens of relations means that the overall number of occurring term pairs that are linked is considered. Types of relations means that the importance of the relation itself is considered independently of how often the term pairs in this relation occur.

The amount of relation tokens might influence the results because very commonly occurring tokens may e.g. not be so informative as less commonly occurring ones. The same holds for types of syntactic relations that may or may not be more informative, independently of whether or not the frequency of their relation is is high or low. Tests in information retrieval have shown that these weights influence retrieval performance significantly (see Salton (18), pp. 204).

Several linguistically motivated tests can be added. One could try to improve the results by using stems instead of lemmatized forms (as does Sparck-Jones (3)). Perhaps one can limit the word classes, i.e. use just terms of one word class for a semantic field. Our goal is to implement our best similarity measure in a connectionist framework. The connectionist model could reduce the small error rates that still remain.

8. Conclusions

In this article we showed that in an area where as yet purely statistically based approaches helped to increase system performance, considerable progress is achieved by the application of natural language processing methods. Early results of methods developed within the TINA project for large textual files look very promising.

References

(1) Salton, G.: On the Use of Term Association in Automatic Information Retrieval. In: Proc. 11th Int. Conf. on Computational Linguistics COLING 86, Bonn 1986. p. 380-386

(2) Giuliano, V: The Interpretation of Word Associations. In: Stevens,E., Giuliano, V. (Eds.): Statistical Association Methods for Mechanized Documentation. Symp.Proc., Washington, DC, March 17-19, 1964. Washington, DC: USGovern.Print.Office 1965. p.25-32. Misc. Publ. 269.

(3) Sparck-Jones,K.: Automatic Keyword Classification for Information Retrieval. London: Butterworth 1971.

(4) Rieger, B.: Semantische Dispositionen. (Semantic Dispositions) In: Rieger,B. (Ed.): Dynamik der Bedeutungskonstitution. (Dynamics of MeaningConstituents). Hamburg: Buske 1984.

(5) Moskovich, W.: Quantitative Linguistics. In: Walker, D., Karlgren, H., Kay, M. (Eds.): Natural Language in Information Science. Fairfax: Band Print 1977. p. 57-74

(6) Lesk, M.: Word-Word Associations in Document Retrieval Systems. Amer. Doc. 20(1969)No.1, p.27-38

(7) Robertson, R., Van Rijsbergen, C., Porter, M.: Probabilistic Models of Indexing and Searching. In: Oddy,R., Robertson,S., Van Rijsbergen,C., Williams,P. (Eds.): Information Retrieval Research. London: Butterworth 1981.

(8) Shaikevich, A.: Automatic Construction of a Thesaurus from Explanatory Dictionaries. Autom. Document. & Mathemat. Linguistics 19(1985)No.2, p.76-89

(9) Chodrow, M., Byrd, R.: Extracting Semantic Hierarchies from a Large On-Line Dictionary. In: Proc. 23rd Annual Meeting of the ACL, Chicago 1985. p.299-304

(10) Das-Gupta, P.: Boolean Interpretation of Conjunctions for Document Retrieval. J. Amer. Soc. Inform. Sci. 38(1987) No.4,p. 245-254

(11) Güntzer, U., Jüttner, G., Seegmüller,G., Sarre, F.: Automatic Thesaurus Construction by Machine Learning from Retrieval Sessions. In: Proc.RIAO 1988. Vol.1, p.587-596

(12) Wettler,M.: A Connectionist System to Simulate Lexical Decisions in Information Retrieval. (to appear in: Pfeifer, R., Schreter, Z., Fogelmann, F., Steels,L.: Connectionism in Perspective. Amsterdam: Elsevier 1989)

(13) Fahlman, S., Hinton,G.: Connectionist Architectures for Artificial Intelligence. Computer (1987)No.1, p.100-109

(14) Schwarz, C.: Content Based Text Handling. In: Inform. Process. & Management" 26(1990)No.2, p.219-226

(15) Schwarz, C.: Automatic Syntactic Analysis of Free-Text. (to appear in: J.Amer.Soc.Inform.Sci. (1990))

(16) Ruge,G.: Strategien zur Entwicklung effizienter Analyseverfahren für die Massentextverarbeitung. (Strategies for the Development of Efficient Analysis of Large Amounts of Textual Data), LDV-Forum 5(1988)No.4, p. 3-11

(17) Sparck-Jones, K., Kay, M.: Linguistics and Information Science. New York 1983.

(18) Salton, G., McGill, M.: Introduction to Modern Information Retrieval. New York, NY: McGraw-Hill 1983.

(19) Ruge, G.: Generierung semantischer Felder auf der Basis von Frei-Texten. (Generation of Semantic Fields on the Basis of Free-Text). LDV-Forum 5(1989)No.2, p.3-17

(20) Panyr, J.: AutomatischeKlassifikation und Information Retrieval. (Automatic Classification and Information Retrieval). Tübingen: Niemeyer 1986.