Hierarchical Orderings of Textual Units

Alexander Mehler

University of Trier
Universitätsring 15
D-54286 Trier, Germany
mehler@uni-trier.de

Abstract

Text representation is a central task for any approach to automatic learning from texts. It requires a format which allows to interrelate texts even if they do not share content words, but deal with similar topics. Furthermore, measuring text similarities raises the question of how to organize the resulting clusters. This paper presents *cohesion trees* (CT) as a data structure for the perspective, hierarchical organization of text corpora. CTs operate on alternative text representation models taking lexical organization, quantitative text characteristics, and text structure into account. It is shown that CTs realize text linkages which are lexically more homogeneous than those produced by minimal spanning trees.

1 Introduction

Text representation is a central task for approaches to text classification or categorization. They require a format which allows to semantically relate words, texts, and thematic categories. The majority of approaches to automatic learning from texts use the vector space or *bag of words model*. Although there is much research for alternative formats, whether phrase- or hyperonym-based, their effects seem to be small (Scott and Matwin, 1999). More seriously (Riloff, 1995) argues that the bag of words model ignores morphological and syntactical information which she found to be essential for solving some categorization tasks. An alternative to the vector space model are *semantic spaces*, which have been proposed as a high-dimensional format for representing relations of semantic proximity. Relying on sparse knowledge resources, they prove to be efficient in cognitive science (Kintsch, 1998; Landauer and Dumais, 1997), computational linguistics (Rieger, 1984; Schütze, 1998), and information retrieval.

Although semantic spaces prove to be an alternative to the vector space model, they leave the question unanswered of how to explore and visualize similarities of signs mapped onto them. In case that texts are represented as points in semantic space, this question refers to the exploration of their implicit, content based relations. Several methods for solving this task have been proposed which range from simple lists via minimal spanning trees to cluster analysis as part of scatter/gahter algorithms (Hearst and Pedersen, 1996). Representing a sign's environment in space by means of lists runs the risk of successively ordering semantically or thematically diverse units. Obviously, lists neglect the poly-hierarchical structure of semantic spaces which may induce divergent thematic progressions starting from the same polysemous unit. Although clustering proves to be an alternative to lists, it seeks a global, possibly nested partition in which clusters represent sets of indistinguishable objects regarding the cluster criterion. In contrast to this, we present *cohesion trees* (CT) as a data structure, in which single objects are hierarchically ordered on the basis of lexical cohesion. CTs, whose field of application is the management of search results in IR, shift the perspective from sets of clustered objects to cohesive paths of interlinked signs.

The paper is organized as follows: the next section presents alternative text representation models as extensions of the semantic space approach. They are used in section (3) as a background of the discussion of cohesion trees. Both types of models, i.e. the text representation models and cohesion trees as a tool for hierarchically traversing semantic spaces, are evaluated in section (4). Finally, section (5) gives some conclusions and prospects future work.

2 Numerical Text Representation

This paper uses semantic spaces as a format for text representation. Although it neglects sentence as well as rhetorical structure, it departs from the bag of words model by referring to paradigmatic similarity as the fundamental feature type: instead of measuring intersections of lexical distributions, texts are interrelated on the basis of the paradigmatic regularities of their constituents. A coordinate value of a feature vector of a sign mapped onto semantic space measures the extent to which this sign (or its constituents in case of texts) shares paradigmatic usage regularities with the word defining the corresponding dimension. Because of this sensitivity to paradigmatics, semantic spaces can capture indirect meaning relations: words can be linked even if they never co-occur, but tend to occur in similar contexts. Furthermore, texts can be linked even if they do not share content words, but deal with similar topics (Landauer and Dumais, 1997). Using this model as a starting point, we go a step further in departing from the bag of words model by taking quantitative characteristics of text structure into account (see below).

Semantic spaces focus on meaning as use as described by the *weak contextual hypothesis* (Miller and Charles, 1991), which says that the similarity of contextual representations of words contributes to their semantic similarity. Regarding the level of texts, reformulating this hypothesis is straightforward:

Contextual hypothesis for texts: the contextual similarity of the lexical constituents of two texts contributes to their semantic similarity.

In other words: the more two texts share semantically similar words, the higher the probability that they deal with similar topics. Clearly, this hypothesis does not imply that texts having contextually similar components to a high degree also share propositional content. It is the structural (connotative), not the propositional (denotative) meaning aspect to which this hypothesis applies. Moreover, this version of the contextual hypothesis neglects the structural dimension of similarity relations: not only that a text is structured into thematic components, each of which may semantically relate to different units, but units similar to the text as a whole do not form isolated, unstructured clumps. Neglecting the former we focus on the latter phenomenon, which demands a supplementary hypothesis:

Structure sensitive contextual hypothesis: units, which are similar to a text according to the contextual hypothesis, contribute to the structuring of its meaning.

Since we seek a model for automatic text representation for which nonlinguistic context is inaccessible, we limit contextual similarity to paradigmatic similarity. On this basis the latter two hypotheses can be summarized as follows:

Definition 1. Let C be a corpus in which we observe paradigmatic regularities of words. The textual *connotation* of a text x with respect to C includes those texts of C, whose constituents realize similar paradigmatic regularities as the lexical constituents of x. The connotation of x is structured on the basis of the same relation of (indirect) paradigmatic similarity interrelating the connoted texts.

In order to model this concept of structured connotation, we use the space model M0 of (Rieger, 1984) as a point of departure and derive three text representation models M1, M2, M3. Since M0 only maps words onto semantic space we extend it in order to derive meaning points of texts. This is done as follows:

M0 analyses word meanings as the result of a two-stage process of unsupervised learning. It builds a lexical semantic space by modeling syntagmatic regularities with a correlation coefficient $\alpha\colon W \to \mathcal{C} \subset \mathbb{R}^n$ and their differences with an Euclidean metric $\delta\colon \mathcal{C} \to \mathcal{S} \subset \mathbb{R}^n$, where W is the set of words, $\mathcal{C}$ is called *corpus space* representing syntagmatic regularities, and $\mathcal{S}$ is called *semantic space* representing paradigmatic regularities. $|W| = n$ is the number of dimensions of both spaces. Neighborhoods of meaning points assigned to words model their semantic similarity: the shorter the points' distances in semantic space, the more paradigmatically similar the words.

The set of words W, spanning the semantic space, is selected on the basis of the criterion of *document frequency*, which proves to be of comparable effectiveness as information gain and χ^2-statistics (Yang and Pedersen, 1997). Fur-

thermore, instead of using explicit stop word lists, we restricted W to the set of lemmatized nouns, verbs, adjectives, and adverbs.

M1: In a second step, we use $\mathcal{S}$ as a format for representing meaning points of texts, which are mapped onto $\mathcal{S}$ with the help of a weighted mean of the meaning points assigned to their lexical constituents:

$$\vec{x}_k = \sum_{a_i \in W(x_k)} w_{ik}\vec{a}_i \in \mathcal{S} \quad (1)$$

$\vec{x}_k$ is the meaning point of text $x_k \in C$, $\vec{a}_i$ the meaning point of word $a_i \in W$, and $W(x_k)$ is the set of all types of all tokens in x_k. Finally, w_{ik} is a weight having the same role as the *tfidf*-scores in IR (Salton and Buckley, 1988). As a result of mapping texts onto $\mathcal{S}$, they can be compared with respect to the paradigmatic similarity of their lexical organization. This is done with the help of a similarity measure σ based on an Euclidean metric δ operating on meaning points and standardized to the unit interval:

$$\sigma : \{\vec{x} \,|\, x \in C\}^2 \rightarrow [0, 1] \quad (2)$$

σ is interpreted as follows: the higher $\sigma(\vec{x}, \vec{y})$ for two texts x and y, the shorter the distance of their meaning points $\vec{x}$ and $\vec{y}$ in semantic space, the more similar the paradigmatic usage regularities of their lexical constituents, and finally the more semantically similar these texts according to the extended contextual hypothesis. This is the point, where semantic spaces depart from the vector space model, since they do not demand that the texts in question share any lexical constituents in order to be similar; the intersection of the sets of their lexical constituents may even be empty.

M2: So far, only lexical features are considered. We depart a step further from the bag of words model by additionally comparing texts with respect to their organization. This is done with the help of a set of quantitative text characteristics used by (Tuldava, 1998) for automatic genre analysis: type-token ratio, hapax legomena, (variation of) mean word frequency, average sentence length, and action coefficient (i.e. the standardized ratio of verbs and adjectives in a text). In order to make these features comparable, they were standardized using z-scores so that random variables were derived with means of 0 and variances of 1. Beyond these characteristics, a further feature was considered: each text was mapped onto a so called *text structure string* representing its division into sections, paragraphs, and sentences as a course approximation of its rhetorical structure. For example, a text structure string

$$(T(D(S))(D(S - S - S))) \quad (3)$$

denotes a text T of two sections D, where the first includes 1 and the second 3 sentences S. Using the Levenshtein metric for string comparison, this allows to measure the rhetorical similarity of texts in a first approximation. The idea is to distinguish units connoted by a text, which in spite of having similar lexical organizations differ texturally. If for example a short commentary connotes two equally similar texts, another commentary and a long report, the commentary should be preferred. Thus, in M2 the textual connotation of a text is not only seen to be structured on the basis of the criterion of similarity of lexical organization, but also by means of genre specific features modeled as quantitative text characteristics. This approach follows (Herdan, 1966), who programmatically asked, whether difference in style correlates with difference in frequency of use of linguistic forms. See (Wolters and Kirsten, 1999) who, following this approach, already used POS frequency as a source for genre classification, a task which goes beyond the scope of the given paper.

On this background a compound text similarity measure can be derived as a linear model:

$$\sigma(x, y) = \sum_{i=1}^{3} \omega_i \sigma_i(x, y) \in [0, 1] \quad (4)$$

a. where $\sigma_1(x, y) = \sigma(\vec{x}, \vec{y})$ models *lexical semantics* of texts x, y according to M1;

b. σ_2 uses the Levenshtein metric for measuring the similarity of the text structure stings assigned to x and y;

c. and σ_3 measures, based on an Euclidean metric, the similarity of texts with respect to the quantitative features enumerated above.

ω_i biases the contribution of these different dimensions of text representation. We yield good results for $\omega_1 = 0.9$, $\omega_2 = \omega_3 = 0.05$.

M3: Finally, we experimented with a text representation model resulting from the aggre-

gation (i.e. weighted mean) of the vector representations of a text in both spaces, i.e. vector and semantic space. This approach, which demands both spaces to have exactly the same dimensions and standardized coordinate values, follows the idea to reduce the noise inherent to both models: whether syntagmatic as in case of vector spaces, or paradigmatic as in case of semantic spaces. We experimented with equal weights of both input vectors.

In the next section we use the text representation models M1, M2, M3 as different starting points for modeling the concept of structured connotation as defined in definition (1):

3 Text Linkage

Departing from ordinary list as well as cluster structures, we model the connotation of a text as a hierarchy, where each node represents a single connoted text (and not a set of texts as in case of agglomerative cluster analysis). In order to narrow down a solution for this task we need a *linguistic criterion*, which bridges between the linguistic knowledge represented in semantic spaces and the task of connotative text linkage. For this purpose we refer to the concept of *lexical cohesion* introduced by (Halliday and Hasan, 1976); see (Morris and Hirst, 1991; Hearst, 1997; Marcu, 2000) who already use this concept for text segmentation. According to this approach, lexical cohesion results from reiterating words, which are semantically related on the basis of (un-)systematic relations (e.g. synonymy or hyponymy). Unsystematic lexical cohesion results from patterns of contextual, paradigmatic similarity: "[...] lexical items having similar patterns of collocation—that is, tending to appear in similar contexts—will generate a cohesive force if they occur in adjacent sentences." (Halliday and Hasan, 1976, p. 286). Several factors influencing this cohesive force are decisive for reconstructing the concept of textual connotation:(i) the contextual *similarity* of the words in question, (ii) their *syntagmatic order*, and (iii) the *distances* of their occurrences. These factors cooperate as follows: the shorter the distance of similar words in a text the higher their cohesive force. Furthermore, preceding lexical choices restrict (the interpretation of) subsequent ones, an effect, which retards as their distance grows. But longer distances may be compensated by higher contextual similarities so that highly related words can contribute to the cohesion of a text span even if they distantly co-occur. By means of restricting contextual to paradigmatic similarity and therefore measuring unsystematic lexical cohesion as a function of paradigmatic regularities, the transfer of this concept to the task of hierarchically modeling textual connotations becomes straightforward. Given a text x, whose connotation is to be represented as a tree T, we demand for any path P starting with root x:

(i) *Similarity:* If text y is more similar to x than z, then the path between x and y is shorter than between x and z, supposed that y and z belong to the same path P.

(ii) *Order:* The shorter the distance between y and z in P, the higher their cohesive force, and vice versa: the longer the path, the higher the probability that the subsequent z is paradigmatically dissimilar to y.

(iii) *Distance:* A cohesive impact is preserved even in case of longer paths, supposed that the textual nodes lying in between are paradigmatically similar to a high degree.

The reason underlying these criteria is the need to control negative effects of intransitive similarity relations: in case that text x is highly similar to y, and y to z, it is not guaranteed that (x, y, z) is a cohesive path, since similarity is not transitive. In order to reduce this risk of incohesive paths, the latter criteria demand that there is a cohesive force even between nodes which are not immediately linked. This demand decreases as the path distance of nodes increases so that topic changes latently controlled by preceding nodes can be realized. In other words: adding text z to the hierarchically structured connotation of x, we do not simply look for an already inserted text y, to which z is most similar, but to a path P, which minimizes the loss of cohesion in the overall tree, when z is attached to P. These comments induce an optimality criterion which tries to optimize cohesion not only of directly linked nodes, but of whole *paths*, thereby reflecting their syntagmatic order. Looking for a mathematical model of this optimality criterion, minimal spanning trees (MST) drop out, since they only optimize direct node-to-node

similarities disregarding any path context. Furthermore, whereas we expect to yield different trees modeling the connotations of different texts, MSTs ignore this *aspect dependency* since they focus on a unique spanning tree of the underlying feature space. Another candidate is given by dependency trees (Rieger, 1984) which are equal to similarity trees (Lin, 1998): for a given root x, the nodes are inserted into its similarity tree (ST) in descending order of their similarity to x, where the predecessor of any node z is chosen to be the node y already inserted, to which z is most similar. Although STs already capture the aspect dependency induced by their varying roots, the path criterion is still not met. Thus, we generalize the concept of a ST to that of a *cohesion tree* as follows:

First, we observe that the construction of STs uses two types of order relations: the first, let it call $\leq_x^1$, determines the order of the nodes inserted dependent on root x; the second, let it call $\leq_y^2$, varies with node y to be inserted and determines its predecessor. Next, in order to build cohesion trees out of this skeleton, we instantiate all relations $\leq_y^2$ in a way, which finds the path of minimal loss of cohesion when y is attached to it. This is done with the help of a distance measure which induces a descending order of cohesion of paths:

Definition 2. Let $G = \langle V, E\rangle$ be a graph and $P = (v_1, \ldots, v_k)$ a simple path in G. The *path sensitive distance* $\hat{\delta}(P, y)$ of $y \in V$ with respect to P is defined as

$$\hat{\delta}(P, y) = \frac{1}{\max(\delta)} \sum_{v_i \in V(P)} \omega_i \delta(\vec{y}, \vec{v}_i) \in [0, 1],$$

where $\sum_{v_i \in V(P)} \omega_i \leq 1$, $\max(\delta)$ is the maximal value assumed by distance measure δ, and $V(P)$ is the set of all nodes of path P.

It is clear that for any of the text representation models M1, M2, M3 and their corresponding similarity measures we get different distance measures $\hat{\delta}$ which can be used to instantiate the order relations $\leq_y^2$ in order to determine the end vertex of the path of minimal loss of cohesion when y is attached to it. In case of increasing biases ω_i for increasing index i in definition (2) the *syntagmatic order* of path P is reflected in the sense that the shorter the distance of x to any vertex in P, the higher the impact of their (dis-)similarity measured by δ, the higher their cohesive force. Using the relations $\leq_y^2$ we can now formalize the concept of a cohesion tree:

Definition 3. Let $G = \langle V, E, \omega\rangle$ be a complete weighted graph induced by a semantic space, and $x \in V$ a node. The graph $D(G, x) = \langle V, \mathcal{E}, \nu\rangle$ with $\mathcal{E} = \{\{v, w\} \mid v <_x^1 w \wedge \neg\exists y \in V : y <_x^1 w \wedge y <_w^2 v\}$ and $\nu \colon \mathcal{E} \to \mathbb{R}$, the restriction of ω to $\mathcal{E}$, is called *cohesion tree* induced by x.

Using this definition of a cohesion tree (CT) we can compute hierarchical models of the connotations of texts, in which not only aspect dependency induced by the corresponding root, but also path cohesion is taken into account.

A note on the relation between CTs and cluster analysis: CTs do not only depart from cluster hierarchies, since their nodes represent single objects, and not sets, but also because they refer to a local, contextsensitive building criterion (with respect to their roots and paths). In contrast to this, cluster analysis tries to find a global partition of the data set. Nevertheless there is a connection between both methods of unsupervised learning: Given a MST, there is a simple procedure to yield a divisive partition (Duda et al., 2001). Moreover, single linkage graphs are based on a comparable criterion as MSTs. Analogously, a given CT can be divided into non-overlapping clusters by deleting those edges whose length is above a certain threshold. This induces, so to say, *perspective clusters* organized dependent on the *perspective of the root and paths* of the underlying CT.

4 Evaluation

Figure (1) exemplifies a CT based on M3 using a textual root dealing with the "BSE Food Scandal" from 1996. The text sample belongs to a corpus of 502 texts of the German newspaper Süddeutsche Zeitung of about 320,000 running words. Each text belongs to an element of a set $\mathcal{T}$ of 18 different subject categories (e.g. politics, sports). Based on the lemmatized corpus a semantic space of 2715 lexical dimensions was built and all texts were mapped onto this space according to the specifications of M3. In figure (1) each textual node of the CT is represented by its headline and subject category as found in the newspaper. All computations

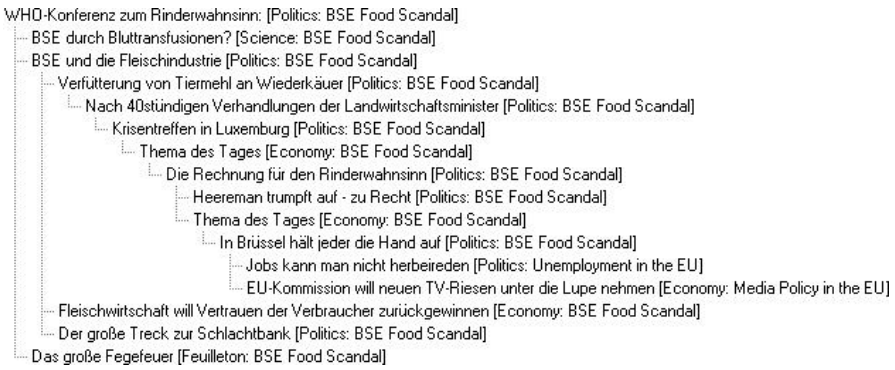

Figure 1: A sample CT.

were performed using a set of C++ programs especially implemented for this study.

In order to rate models M1, M2, M3 in comparison to the vector space model (VS) using MSTs, STs and CTs as alternative hierarchical models we proceed as follows: as a simple measure of representational goodness we compute the average categorial cohesion of links of all MSTs, STs and CTs for the different models and all texts in the corpus. Let $G = \langle V, E \rangle$ be a tree of textual nodes $x \in V$, each of which is assigned to a subject category $\tau(x) \in \mathcal{T}$, and $P(G)$ the set of all paths in G starting with root x and ending with a leaf, then the categorial cohesion of G is the average number of links $(v_i, v_j) \in E$ per path $P \in P(G)$, where $\tau(v_i) = \tau(v_j)$. The more nodes of identical categories are linked in paths in G, the more categorially homogeneous these paths, the higher the average categorial cohesion of G. According to the conceptual basis of CTs we expect these trees to be of highest categorial link cohesion, but this is not true: MSTs produce the highest cohesion values in case of VS and M3. Furthermore, we observe that model M3 induces trees of highest cohesion and lowest variance, whereas VS shows the highest variance and lowest cohesion scores in case of STs and CTs. In other words: based on semantic spaces, models M1, M2, and M3 produce more stable results than the vector space model.

Using M3 as a starting point it can be asked more precisely, which tree class produces the most cohesive model of text connotation. Clearly, the measure of categorial link cohesion is not sufficient to evaluate the classes, since two immediately linked texts belonging to the same subject category may nevertheless deal with different topics. Thus we need a finer-grained measure which operates directly on the texts' meaning representations. In case of unsupervised clustering, where fine-grained class labels are missed, (Steinbach et al., 2000) propose a measure which estimates the overall cohesion of a cluster. This measure can be directly applied to trees: let $P_{v_1,v_n} = (v_1, \ldots, v_n)$ be a path in tree $G = \langle V, E \rangle$ starting with root $v_1 = x$, we compute the cohesion of P *irrespective of the order of its nodes* as follows:

$$\xi(P_{v_1,v_n}) = 1 - \frac{1}{n^2} \sum_{i,j=1}^{n} \frac{1}{\max(\delta)} \delta(v_i, v_j) \qquad (5)$$

The more similar the nodes of path P according to metric δ, the more cohesive P. δ is derived from the distance measure operating on the semantic space to which texts v_i are mapped. As before, all scores $\xi(P)$ are summed up for all paths in $P(G)$ and standardized by means of $|P(G)|$. This guarantees that neither trees of maximum height (MHT) nor of maximum degree (MDT), i.e. trees which trivially correspond to lists, are assigned highest cohesion values. The results of summing up these scores for all trees of a given class for all texts in the test corpus are shown in table (2). Now,

Model	MSTs	STs	CTs
VS	1325.88	462.04	598.87
M1	1093.06	680.06	1185.92
M2	1097.39	661.72	1168.63
M3	1488.38	628.51	1032.55

Table 1: Alternative representation models and scores of trees derived from them.

Type	$\sum \xi(G)$	Type	$\sum \xi(G)$
MDT	388.1	MST	416.3
MHT	388.1	DT	430.9
RST	386.6	CT	438.6

Table 2: The sum of the cohesion scores for all tree classes and all texts in the test corpus.

CTs and STs realize the most cohesive structures. This is more obvious if the scores $\xi(G)$ are compared for each text in separation: in 494 cases, CTs are of highest cohesion according to measure (5). In only 7 cases, MST are of highest cohesion, and in only one case, the corresponding ST is of highest cohesion. Moreover, even the stochastically organized so called *random successor trees* (RST), in which successor node's and their predecessors are randomly chosen, produce more cohesive structures than lists (i.e. MDTs and MHTs), which form the predominant format used to organize search results in Internet.

To sum up: Table (2) rates CTs in combination with model M3 on highest level. Thus, from the point of view of lexical semantics CTs realize more cohesive branches than MSTs. But whether these differences are significant, is hard to evaluate, since their theoretical distribution is unknown. Thus, future work will be on finding these distributions.

5 Conclusion

This paper proposed 3 numerical representation formats as means for modeling the hierarchical connotation of texts in combination with cohesion trees. This was done by extending the weak contextual hypothesis onto the level of texts in combination with a reinterpretation of the concept of lexical cohesion as a source for text linkage. Although the formats used depart from the bag of words model there is still the need of investigating numerical formats which rely on linguistically more profound discourse models.

References

R. O. Duda, P. E. Hart, and D. G. Stork. 2001. *Pattern Classification*. Wiley, New York.

Michael A. K. Halliday and R. Hasan. 1976. *Cohesion in English*. Longman, London.

M. A. Hearst and J. O. Pedersen. 1996. Reexamining the cluster hypothesis: Scatter/gather on retrieval results. In *Proc. ACM SIGIR*.

M. A. Hearst. 1997. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64.

G. Herdan. 1966. *The Advanced Theory of Language as Choice and Chance*. Springer, Berlin.

W. Kintsch. 1998. *Comprehension. A Paradigm for Cognition*. Cambridge University Press.

T. K. Landauer and S. T. Dumais. 1997. A solution to plato's problem. *Psychological Review*, 104(2):211–240.

D. Lin. 1998. Automatic retrieval and clustering of similar words. In *Proc. COLING-ACL*.

D. Marcu. 2000. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, Cambridge, Massachusetts.

G. A. Miller and W. G. Charles. 1991. Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

J. Morris and G. Hirst. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1):21–48.

B. Rieger. 1984. Semantic relevance and aspect dependency in a given subject domain. In *Proc. 10th COLING*.

E. Riloff. 1995. Little words can make a big difference for text classification. In *Proc. SIGIR-95*.

G. Salton and C. Buckley. 1988. Term weighting approaches in automatic text retrieval. *Information Processing Management*, 24(5):513–523.

H. Schütze. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

S. Scott and S. Matwin. 1999. Feature engineering for text classification. In *Proc. 16th ICML*, pages 379–388.

M. Steinbach, G. Karypis, and V. Kumar. 2000. A comparison of document clustering techniques. In *KDD Workshop on Text Mining*.

J. Tuldava. 1998. *Probleme und Methoden der quantitativ-systemischen Lexikologie*. Wissenschaftlicher Verlag, Trier.

M. Wolters and M. Kirsten. 1999. Exploring the use of linguistic features in domain and genre classication. In *Proc. EACL*.

Y. Yang and J. O. Pedersen. 1997. A comparative study on feature selection in text categorization. In *Proc. 14th ICML*.