INTERNATIONAL DESIGN CONFERENCE - DESIGN 2006
Dubrovnik - Croatia, May 15 - 18, 2006.

DESIGN 2006

SEMANTIC ANNOTATION TO SUPPORT AUTOMATIC TAXONOMY CLASSIFICATION

S. Kim, R.H. Bracewell, S Ahmed and K.M. Wallace

Keywords: corporate taxonomy classification, semantic annotation, natural language processing, rhetorical structural theory

1. Introduction

Electronic documents are used throughout a product lifecycle and are created and accessed by users working in substantially different business operations and with very different areas of expertise. One way to organise corporate documents into easily accessible formats is to use a corporate taxonomy. A corporate taxonomy is the hierarchical organisation of the concepts available within an enterprise. It provides a holistic view of the organisation's information architecture leading to better information flow between business operations and improved accessibility to unstructured texts. Unlike classic taxonomies (e.g. medical, biology, etc) in which objects map onto just one entry, a corporate taxonomy needs to be flexible in order to encompass diverse business operations. It must also facilitate searching and navigating through the multiple facets of enterprises. A corporate taxonomy should allow texts to be encoded with multiple metadata reflecting each operation's domain model. However, organisations still need to maintain an integrated view of their documents. One way to maintain a balance between ensuring common access to the taxonomy and dealing with diverse business interests of the users is to focus on the main messages that the authors of the documents intended to communicate. These messages contain the information that people are most likely to reuse or refer to in the future. If these main messages can be mapped onto the taxonomy, the goals of consistent access to the taxonomy and maximum information sharing might be achieved.
In organisations, people create documents to distribute information rather than to convey the views of individuals or groups. A document is encoded with various semantics and accessed by users who have very different interests. For example in the case of product reviews by customers, negative and positive customer opinions are the main messages for market researchers who are trying to assess market demands. On the other hand, designers are more interested in design-related issues, comments and problems. It would therefore be beneficial to include those semantics that facilitate searching for information in a way that reflects the interests of the users. For example, for a designer whose task is to reduce fan noise, guidance on how to minimise aerodynamic noise should be retrieved. On the other hand, if that designer is more interested in using a specific method for noise reduction, then documents describing the methods along with their advantages or disadvantages are more useful. It is therefore important to investigate how to extract and use the most appropriate semantics for indexing and searching documents using a corporate taxonomy. In general, taxonomy development proceeds by defining initial taxonomy entities, arranging them into a category tree and assigning information to each category. For the first two of those stages, manual intervention is necessary and professional indexers are often used in the process. However, when a taxonomy is to be used for indexing and retrieving documents, it is not cost effective nor practical to rely solely on the professionals for cataloguing corporate documents. Domain expertise is essential for managing corporate information and it is hard to find indexers who are familiar with both the taxonomy and the domain models. It is

also difficult to transfer the required domain knowledge to others when experts retire or move to other departments. Some organisations, therefore, prefer to use automatic classifications.

Current automatic classifications have an accuracy of between 50-80%, meaning that up to half the documents are missing or misclassified. There are many reasons for this, including insufficient training data and noise in labelled documents. This paper focuses on two main observations. The first observation is referred to as "term-mismatch", which means that people might use different terms when describing similar concepts. Because of the inherent ambiguity in natural language texts, automatic term expansion is still not perfect and manual intervention often necessary. However, it is not necessary to create additional related terms for every term in the texts, since too many indexing terms can increase complexity and decrease accuracy. It is known that not all terms are equally important in order to understand the underlying concepts. A method that clearly identifies those terms and sentences that are directly related to the classification task is therefore needed. The second observation relates to the different ways individuals chose to index documents and how automatic classification is implemented. When people index a document manually, i.e. by subject-matter type, they tend to make classification decisions based on a small number of *meaningful* sentences. Experts intuitively recognise those parts of the texts that are key to understanding the main messages that the authors intended to convey. These terms and sentences contain the information that people are likely to want to reuse or refer to in the future.

Term-based indexing is common in current automatic classifications. A term weight is assigned according to a term's frequency within a document and multiplied by the inverse document frequency. In this way, sentences are treated equally in such a way that a text is summarised as a list of the most frequent terms drawn from the entire content. Some terms might result in noisy classification criteria. As such, less-important or irrelevant terms should be removed from the classification rules. Most approaches select terms based on their number of occurrences. However, some sentences are more essential in order to understand the contents than others, and terms from *important* sentences should be identified in the classification process. If were feasible to deduce what terms trigger people to classify a given text into certain categories, classification could be improved. This paper presents a new taxonomy classification method that generates classification criteria from a small number of important sentences identified through semantic annotations, e.g. cause-effect. Rhetorical Structure Theory (RST) is used to discover the semantics (Mann *et al.* 1988). Specifically, the annotations identify which parts of a text are more important for understanding its contents. The extraction of salient sentences is a major issue in text summarisation. Commonly used methods are based on statistical analysis, but for subject-matter type texts, linguistically motivated natural language processing techniques, like semantic annotations, are preferred. An experiment to test the method using 140 documents collected from industry demonstrated that classification accuracy can be improved by up to 16%.

2. Literature Review

There are three main approaches to taxonomy classification. The first approach, manual classification, requires domain experts to assign documents to categories. It often outperforms the other two approaches but requires significant investment in training cost, and retraining is required when the taxonomy changes. The second approach, automatic (supervised) classification, learns the basis for classification by extracting common features from labelled documents. Statistical machine learning methods, e.g. naïve Bayesian classifier, are commonly used (Mitchell 1997). One way to represent documents is to use the term-weights specifying the numeric contribution of a given term to predict the correct categories. With this representation, classifiers deduce the profile of each category, which is the summary of the classified documents. These profiles are used to generate a similarity value between an unlabelled document and each category. Classifiers perform well if labelled documents and unlabelled texts share a sufficient number of terms due to similar styles and vocabularies. For some organisations, it is rather difficult to prepare a large number of training examples that the algorithms depend on. The third approach, semi-automatic classification, actively involves users in the training step to help the classifier identify the most informative documents to label and to use for training (Seung *et al.* 1992). It can significantly reduce the number of training examples needed. In

addition, users can improve the classifications by accepting or rejecting recommended suggestions so that the classifiers can update their classification rules. Most efforts to identify important sentences are based on text summarisation, which extracts representative information from original contents thus offering more manageable and easily transmitted formats (Mani *et al.* 1999). Ko *et al.* (Ko *et al.* 2002) used a statistical-based summarisation to assign high weights to the terms that occurred in important sentences for text categorisation. The method proposed in this paper, on the other hand, relies on linguistic features to measure the importance of sentences and it is hoped that this will improve taxonomy classification performance.

3. Background to the Method

As an example of engineering taxonomy, Engineering Design Integrated Taxonomy (EDIT) is used throughout this paper (Ahmed 2005). EDIT consists of four root concepts: (1) process, (2) product, (3) function, and (4) issue. This paper focuses on the classification of the *issue* root concept. According to Ahmed (Ahmed 2005), issues are considerations designers must take into account when carrying out a design process. These can be descriptions of problems arising during a product's lifecycle or new design requirements to be satisfied. In comparison to the *product* root class, which is a hierarchical list of product names, issues are collections of engineering design topics.

In order to identify what techniques should be used for the classifying issues, a corpus analysis of sample documents was carried out. One of the authors examined 140 problem reports obtained from a large engineering company that describe problems that arose during product development. The corpus analysis identified the characteristic content of the documents from both a structural property perspective and a semantic perspective. It revealed that rather than just reporting problems other semantics were present in the documents. The reason for this is likely to be because engineers have very individual interpretations of what constitutes a *problem*. The linguistic analysis showed that engineers use the problem reports: (1) to report problems, (2) to suggest potential improvement, (3) to report general requirements, and (4) to summarise changes made. Another of the authors manually indexed the same documents against the *issue* root concept in EDIT, and observed that:

- it is difficult identify issues by simply looking up the *issue* terms
- there are some key sentences that are particularly relevant to issues.

The first observation highlights the need for a thesaurus. The list issues was derived mainly through observational studies, e.g. interviewing designers to extract the issue hierarchy. The aim was to provide more intuitive navigation and search structures for engineering designers by identifying the issues that most closely reflected how designers actually described specific topics. For example, engineers seldom use the exact term *dynamic response*, which is an entity in EDIT, in documents when describing problems caused by the distribution of stiffness in product structures. Instead, they often describe the causes that triggered the problem. This means that that an automatic classification method needs to access the conceptual descriptions of the issues. Term expansion is domain-dependent and difficult to fully automate. One problem is that a thesaurus does not remain static as new expressions and abbreviations arise frequently. It is also not wise to create additional related terms for every term since too many indexing terms can increase complexity and decrease accuracy. It is therefore essential for organisations to identify which domain concepts need to be accessed by diverse vocabularies and should be extended by adding related terms. The second observation highlights the fact that technical reports are free-text formats, i.e. engineers are free to enter contents of any length, style and content. As the documents are meant to be shared by other engineers, background or introductory statements are often included to clarify the observations of the authors. Some engineers elaborate the problems with their experience of previous similar problems or product situations. Some users find this additional information useful as it can increase their confidence in the claims made by the authors. However, when the engineers index the texts against the taxonomy, they seldom make category decisions based on this background information. Instead, the messages extracted from a small number of *important* sentences dominate, since it is these messages that the engineers most need to reuse or refer to in the future.

Both observations highlight the need to differentiate important sentences from unimportant ones, and to favouring the former when extracting the underlying contents. The method being developed uses

the semantic relations between the sentences in order to provide an improved content analysis. A new classifier is proposed that improves accuracy through a more sophisticated indexing.

4. Description of the Method

Supervised classification methods rely on training examples to learn classification criteria. Typical approaches use terms extracted from the entire contents and this can result in noisy classification rules. The proposed method creates a classifier from a small number of important sentences determined from semantic annotations. Figure 1 shows the overall procedure of the method. Currently, semantic annotations are created manually.

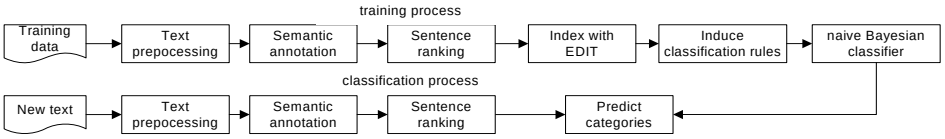

Figure 1. Overview of the procedure

4.1 Text Preprocessing

Natural language processing can improve a current term-based indexing by representing terms, e.g. cook, not only with their frequencies but also based on their linguistic features, i.e. syntactic roles (noun) or meanings (someone who cooks). With this linguistic information, it is easier to expand the terms with interchangeable words, e.g. chef. A text is first analysed using Apple Pie parser (Sekine 2001) for part-of-speech taggings and subsequently very common words, e.g. 'a' and 'the', are deleted.

4.2 Semantic Annotations

A small number of *meaningful* sentences are essential to understand the contents of texts and these contribute to improved taxonomy classification. Most efforts to identifying important sentences draw on a text summarisation that extracts representative information from original contents. These sentences occur in different positions in a text, which means that the structural positions are not very helpful. Commonly used methods are based on statistical analysis, but for subject-matter type texts linguistically motivated natural language processing techniques, like semantic annotations, are preferred. Semantic relations are the relationships between two text spans in a text. Current annotation efforts, particularly within semantic Web community, centre on the automatic extraction of named-entities, including the associations between two named-entities in a single sentence. In the proposed method, annotations are based on computational linguistic theory, i.e. Rhetorical Structure Theory (RST) analysis, which defines a set of rhetorical relations and use them to describe how the sentences are combined to form a coherent text (Mann *et al.* 1988). As such, RST analysis discovers relations within a sentence or among sentences. Since sentences are not properly comprehensible when isolated, this approach provides a more sophisticated content analysis. A total of 30 relations are defined in RST but the corpus analysis with the sample texts indicates only a subset of them, i.e. *background, cause-effect, condition, contrast, elaboration, evaluation, joint, means, purpose, solutionhood and summary* are needed in this paper.

In RST, the two text spans are further differentiated as *nucleus* and *satellite*. The nucleus texts are more essential to the overall purpose of the document and are comprehensible independently of the satellite. For example, consider the following two sentences: *(1)This flight takes 5½ hours*; *(2)So there's a stop-over in Paris*. An *evidence* relation is identified, with sentence (2) being the nucleus since sentence (1) is only used to increase reader's belief in the author's claim that the plane will land in Paris. *Contrast* and *joint* are multinuclear types, which means that no particular spans are more central to the communication.

Figure 2 shows RST analysis using an example text (Marcu 1999). The text shown in diagram (a) is decomposed into elementary discourse units (those surrounded by square brackets) and diagram (b) shows the corresponding RST analysis. Nuclei are linked by solid lines and satellites by dotted lines.

The leaves of the tree correspond to elementary discourse units and the internal nodes correspond to contiguous text spans. For example, a *justification* relation holds between the nucleus labelled as 2 and the satellite labelled as 1.

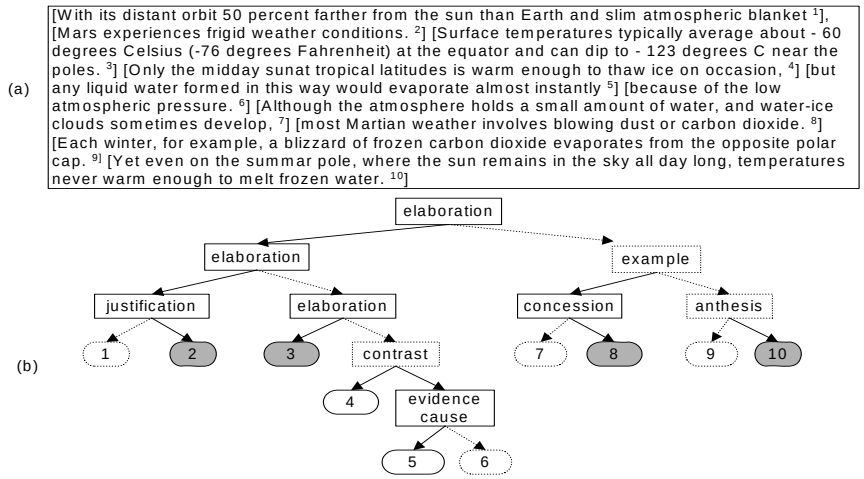

Figure 2. An example of RST analysis (Marcu 1999)

Various linguistic features can signal the relations, e.g. pronouns and other referents, cue phrases or syntactic similarities. Using pre-defined cue phrases is common and easy to implement. Although the low presence of such phrases can lead to many undiscovered relations, they can serve as a reference for annotators. A text is decomposed into a set of text spans (nucleus -satellite pairs) with which coherence relations are associated. Table 1 summarises cue phrases extracted from Knott (Knott *et al.* 1992) and Williams (Williams *et al.* 2003).

Table 1. Cue phrases for identifying semantics

Semantics	Cue phrases
Background	With, probably
Cause-Effect	Because, since, as, as a consequence, as a result, thus, therefore, due to, lead to, consequently
Condition	as long as, if…then, if, so long as, unless, until
Contrast	although, by contrast, even though, however, though, whereas, while
Elaboration	also, in addition, in particular, for example, in general
Evaluation	with, so, but, which, even so
Joint	and
Means	by, with, using
Purpose	in order to, for the purpose of
Solutionhood	proposed solution, options
Summary	that is, in other words, in short, to summarize, summarising

4.3 Indexing by the Importance of Sentences

With RST analysis, it is possible to induce a partial ordering on the importance of all the text spans by using nucleus text spans and the tree-based rhetorical structure of the text (Marcu 1999). As shown in Figure 2, a text is represented as a binary tree in RST. The idea is that a nucleus node (described in solid line) gets a higher importance weighting than a satellite span (described in dotted line). This recursively applied to an overall discourse tree. In addition, a nucleus that is closer to the top node of a discourse tree is assumed to be more important than the nucleus texts at the bottom. In Figure 2(b), unit 2 is the most important sentence, followed by units 8, 3 and 10. Unit 3 gets a lower importance

weighting than unit 2 since the latter is the nucleus of the root node. The importance score, $s(u_k, D, l)$, of a sentence unit (u_k), in a discourse tree (D) for a given document that has depth (l), is determined by the following conditions (Marcu 1999):

- If u_k is among the nucleus of a node which has a depth, l, then its score is l
- Otherwise, returns $max(s(u_k, C(D), l-1))$, in which $C(D)$ is the child sub-trees of each node.

The discourse tree shown in Figure 2 has depth of 6 (i.e. the number of nodes along the longest path from the root node down to the farthest leaf node). Using the first condition above, discourse unit 2 gets a score of 6 since it is one of the nucleus units of the root node, which has a depth 6; whereas unit 3 gets a score of 4, since it is the nucleus span of a node that is located two levels below the root. These scores are used to select the top 'n' most important sentences in a given text. The selected sentences are then used for creating classification criteria by the naive Bayesian classifier, which is one of the supervised methods. It is based on a simplified Bayesian theorem that assumes that terms are independent in class (Mitchell 1997). $P(c_m \mid d_j)$ denotes the probability that taxonomy index, c_m, will be a category which the document, d_j, (represented as $t_1..t_i$) would be sorted into and it is defined as:

$$P(c_m \mid d_j) = P(c_m) \prod_{p \in positions} P(t_i \mid c_m) \quad (1)$$

$P(t_i \mid c_m) = \frac{(n_i + 1)}{(N + |vocabulary|)}$, where n_i is the number of times the term, t_i, occurs in the taxonomy class c_m, N is the total number of terms in class c_m, *vocabulary* is the set of all distinct terms in all taxonomy classes, |*vocabulary*| is the total number of distinct terms in all taxonomy indexes, *positions* is the set of terms that appear both in document, d_j, and in *vocabulary*.

$$P(c_m) = \frac{number\ of\ documents\ in\ c_m}{total\ number\ of\ documents\ in\ across\ classes}$$

The naïve Bayesian computes similarities between a new text and all taxonomy classes using equation (1), and then compares the values to generate the maximal probability.

5. Testing the Method

An experiment was undertaken to test whether RST-based annotations are efficient in differentiating important sentences from unimportant ones and hence improved classification. One of the authors examined the documents, annotated them with the semantics defined in section 4.2. RSTTool was used for the annotations (O'Donnell 2000). It provides a graphical interface with which users can manually annotate the rhetorical relations. Cue phrases in Table 1 were used to segment text spans and select a relation, but if these were not acceptable, the annotator chose others. A total of 957 semantics were annotated for 140 technical reports. Table 2 summarizes the annotation results including the total number of occurrence (68) and the unique number of documents (56) in which each semantic, e.g. *Background*, occurred. *Elaboration* (17%) and *Joint* (23%) were the most common and this finding is consistent with the results obtained by Williams (Williams *et al.* 2003). However, the frequent occurrences of *Purpose* (14%) and *Cause-Effect* (13%) were rather surprising since by considering the semantics of the dataset *Solutionhood* was expected to be the most frequent. In fact, *Solutionhood* was rather rare. A close examination revealed that some documents only reported problems without mentioning solutions. Since most semantics in RST are binary, if either a solution or a problem is omitted, it is missed. This examination also confirmed that because the reports were written by individuals, each of whom has different ways of defining problems, the dataset contained other types than simply problem issues. A total of 67, 46, 24, and 3 documents were identified as types of *problem issue*, *potential improvement*, *general requirements* and *changes made*, respectively. Table 2 also

shows the proportion of each semantic to the total number of classified documents for each document type. For example, approximately 31% of the indexed documents which are *problem issue* type have *Background* semantic. It also shows a strong correlation between a *problem issue* type and *Solutionhood* and *Cause-Effect* annotations. *Purpose, Cause-Effect* and *Means* were used frequently to suggest potential improvement.

Table 2. The distribution of annotations in the corpus

Semantics	No. of occurrences	No. of documents	Document types			
			Problem issue	Potential improvement	General requirement	Changes made
Background	68	56	31%	28%	46%	33%
Cause-Effect	121	82	60%	33%	38%	33%
Condition	24	20	12%	20%	4%	0
Contrast	45	35	24%	13%	17%	0
Elaboration	161	90	60%	50%	46%	67%
Evaluation	54	45	28%	24%	20%	33%
Joint	218	97	60%	52%	46%	67%
Means	30	28	12%	30%	0	0
Purpose	134	96	52%	50%	63%	67%
Solutionhood	75	64	58%	15%	17%	0
Summary	27	30	10%	24%	25%	0

A total of 193 indexes were used for the 140 technical reports. Each text was analysed according to the procedure described in Section 4. Equation (1) was used for creating the classifying criteria. For a comparison, the naïve Bayesian classifier based on the entire contents was constructed. Table 3 shows the result. The proposed classifier, with a maximum 65% accuracy, outperformed the classification based on full texts (49%). Moreover, six more categories were correctly predicted compared to that of the full-length classifier. That is, the proposed classifier made correct classifications for six categories that were not extracted by using the basic method. It also shows that using six most important sentences based on the semantic annotations is beneficial and too few sentences actually decreases the accuracy. That is, the classification accuracy by using 1-2 sentences is lower than the accuracy by using full contents.

Table 3. Naïve Bayesian classification result

	Number of top 'n' most important sentences used for classification										Full text
	1	2	3	4	5	6	7	8	9	10	
Accuracy	37%	41%	49%	55%	62%	65%	64%	60%	58%	54%	49%

6. Conclusions and Future Work

The recognition of important sentences is clearly useful for a text classification. The idea developed is that each sentence in a document makes a different contribution to understanding the contents. The terms extracted from the more important sentences are the most useful for classifying the texts. The proposed method, with 65% accuracy, outperformed an indexing method that used full-length contents. The experimental results also indicated that semantic annotations derived from RST analysis were an efficient way to distinguish important sentences from unimportant ones.

Whereas it is true that a taxonomy enhances information searches by providing navigational structures, some users still find it difficult to locate the information they need. Many Intranet-based searches (approximately 60-80%) are driven by recall, i.e. users look for documents that they have seen before. Users may recall some attributes (date-of-creation, author, etc) or semantics (problem, causes, etc), but are not able to specify exact indexed keywords. Some commercial search engines

already allow users to narrow down their queries with such attributes, but semantic-based searches have not yet been explored. It is more difficult to employ semantics since computers search for keywords, whereas users explore concepts. In addition, there are numerous ways to describe concepts, e.g. causal relations can be expressed using conjunctives (because), adverbs (due to, leading to), verb (cause, make) and so on. The annotations proposed in this paper can abstract users' queries into one of the defined semantics thus enabling concept-based searches.
In organisations, relevant information is often scattered over different documents so that users spend considerable time searching for and integrating information in a meaningful way. With well-defined semantics, it is possible to extract dependence associations between sentences (cause-effect), redundant sentences (restatement), and contradictory claims (contrast). Thereby the annotations proposed here could increase document accessibility. The proposed annotation could also improve the user interface. Current search engines answer users' queries with a list of ranked documents that might mention answers to the queries. When users look for specific answers, they have to sift through the retrieved texts and examine the contents. With the proposed semantics, users could express their information needs in natural language queries (e.g. *What are the major causes for aircraft crashes?*) and receive concise answers (e.g. *pilot error, mechanical failure*) that have been extracted through semantic relations (e.g. cause-effect).
Finally, it is very labour-intensive to manually annotate texts with semantics, even assisted by a software tool (e.g. RSTTool). It took two weeks for one of authors to create the testing corpus (see Section 4.2). The automatic discovery of the annotations, especially from unstructured texts, has not been explored fully. This will be the focus of future work.

Acknowledgement

This work was funded by the University Technology Partnership for Design, which is a collaboration between Rolls-Royce, BAE SYSTEMS and the Universities of Cambridge, Sheffield and Southampton.

References

Ahmed, S., "Encouraging reuse of design knowledge: a method to index knowledge", Design Studies, Vol. 26, No. 6, 2005, pp. 565-592.

Knott, A., Dale, R., "Using linguistic phenomena to motivate a set of rhetorical relations", Technical report HCRC/RP-30, University of Edinburgh,1992

Ko, Y., Park, J., Seo, J., "Automatic Text Categorization using the Importance of Sentences",Proceedings of the 19th International Conference on Computational Linguistics,2002,

Mani, I., Maybury, M., "Advances in Automatic Text Summarisation", The MIT Press, 1999

Mann, W., Thompson, S., "Rhetorical structure theory: Toward a functionaltheory of text organization", Text, Vol.8, No. 3,1988, pp. 243-281

Marcu, D., "Discourse trees are good indicators of importance in text", Advances in Automatic Text Summarization, Mani, I., Maybury, M. (eds.), MIT Press, 1999

Mitchell, T. M., "Machine Learning", McGraw-Hill International Editions, 1997

O'Donnell, M., "RSTTool 2.4 -- A Markup Tool for Rhetorical Structure Theory", Proceedings of the International Natural Language Generation Conference (INLG'2000), 2000, pp.253 - 256

Sekine S., Grishman R, "A Corpus-Based Probabilistic Grammar with only two Non-Terminals", 1st International Workshop on Multimedia Annotation, 2001

Seung, H. S., Opper, M., Sompolinsky, H, "Query by committee", Proceedings of the fifth annual workshop on Computational Learning Theory", 1992, pp.287-294

Williams, S., Reiter, E., "A corpus analysis of discourse relations for Natural Language Generation", Proceedings of Corpus Linguistics,2003

Dr. Sanghee Kim
Engineering Design Centre, Department of Engineering, University of Cambridge
Trumpington Street, Cambridge, CB2 1PZ, UK
Tel.: +44 1223 760559
Fax.: +44 1223 332662,
Email: shk32@eng.cam.ac.uk