Using decision trees to select the grammatical relation of a noun phrase

Simon CORSTON-OLIVER
Microsoft Research
One Microsoft Way
Redmond WA 98052, USA
simonco@microsoft.com

Abstract

We present a machine-learning approach to modeling the distribution of noun phrases (NPs) within clauses with respect to a fine-grained taxonomy of grammatical relations. We demonstrate that a cluster of superficial linguistic features can function as a proxy for more abstract discourse features that are not observable using state-of-the-art natural language processing. The models constructed for actual texts can be used to select among alternative linguistic expressions of the same propositional content when generating discourse.

1. Introduction

Natural language generation involves a number of processes ranging from planning the content to be expressed through making encoding decisions involving syntax, the lexicon and morphology. The present study concerns decisions made about the form and distribution of each "mention" of a discourse entity: should reference be made with a lexical NP, a pronominal NP or a zero anaphor (i.e. an elided mention)? Should a given mention be expressed as the subject of its clause or in some other grammatical relation?

If all works well, a natural language generation system may end up proposing a number of possible well-formed expressions of the same propositional content. Although these possible formulations would all be judged to be valid sentences of the target language, it is not the case that they are all equally likely to occur.

Research in the area of Preferred Argument Structure (Corston 1996, Du Bois 1987) has established that in discourse in many languages, including English, NPs are distributed across grammatical relations in statistically significant ways. For example, transitive clauses tend not to contain lexical NPs in both subject and object positions and subjects of transitives tend not to be lexical NPs nor to be discourse-new.

Unfortunately, the models used in PAS have involved only simple chi-squared tests to identify statistically significant patterns in the distribution of NPs with respect to pairs of features (e.g. part of speech and grammatical relation). A further problem from the point of view of computational discourse analysis is that many of the features used in empirical studies are not observable in texts using state-of-the art natural language processing. Such non-observable features include animacy, the information status of a referent, and the identification of the gender of a referent based on world knowledge.

In the present study, we treat the task of determining the appropriate distribution of mentions in text as a machine learning classification problem: what is the probability that a mention will have a certain grammatical relation given a rich set of linguistic features? In particular, how accurately can we select appropriate grammatical relations using only superficial linguistic features?

2. Data

A total of 5,252 mentions were annotated from the Encarta electronic encyclopedia and 4,937 mentions from the Wall Street Journal (WSJ). Sentences were parsed using the Microsoft English Grammar (Heidorn 1999) to extract mentions and linguistic features. These analyses were then hand-corrected to eliminate noise in the training data caused by inaccurate parses, allowing us to determine the upper bound on accuracy for the classification task if the computational analysis were perfect. Zero anaphors were annotated only when they occurred as subjects of coordinated clauses. They have been excluded

from the present study since they are invariably discourse-given subjects.

3. Features

Nineteen linguistic features were annotated, along with information about the referent of each mention. On the basis of the reference information we extracted the feature [InformationStatus], distinguishing "discourse-new" versus "discourse-old". All mentions without a prior coreferential mention in the text were classified as discourse-new, even if they would not traditionally be considered referential. [InformationStatus] is not directly observable since it requires the analyst to make decisions about the referent of a mention.

In addition to the feature [InformationStatus], the following eighteen observable features were annotated. These are all features that we can reasonably expect syntactic parsers to extract with sufficient accuracy today or in the near future.

- [ClausalStatus]: Does the mention occur in a main clause ("M"), complement clause ("C"), or subordinate clause ("S")?
- [Coordinated] The mention is coordinated with at least one sibling.
- [Definite] The mention is marked with the definite article or a demonstrative pronoun.
- [Fem] The mention is unambiguously feminine.
- [GrRel] The grammatical relation of the mention (see below, this section).
- [HasPossessive] Modified by a possessive pronoun or a possessive NP with the clitic *'s* or *s'*.
- [HasPP] Contains a postmodifying pre-positional phrase.
- [HasRelCl] Contains a postmodifying relative clause.
- [InQuotes] The mention occurs in quoted material.
- [Lex] The specific inflected form of a pronoun, e.g. *he*, *him*.
- [Masc] The mention is unambiguously masculine.
- [NounClass] We distinguish common nouns versus proper names. Within proper names, we distinguish the name of a place ("Geo") versus other proper names ("ProperName").
- [Plural] The head of the mention is morphologically marked as plural.
- [POS] The part of speech of the head of the mention.
- [Prep] The governing preposition, if any.
- [RelCl] The mention is a child of a relative clause.
- [TopLevel] The mention is not embedded within another mention.
- [Words] The total number of words in the mention, discretized to the following values: {0, 1, 2, 3, 4, 5, 6to10, 11to15, above15}.

Gender ([Fem], [Masc]) was only annotated for common nouns whose default word sense is gendered (e.g. "mother", "father"), for common nouns with specific morphology (e.g. with the *–ess* suffix) and for gender-marked proper names (e.g. "John", "Mary"). Gender was not marked for pronouns, to avoid difficult encoding decisions such as the use of generic "he". [1] Gender was also not marked for cases that would require world knowledge.

The feature [GrRel] was given a much finer-grained analysis than is usual in computational linguistics. Studies in PAS have demonstrated the need to distinguish finer-grained categories than the traditional grammatical relations of English grammar ("subject", "object" etc) in order to account for distributional phenomena in discourse. For example, subjects of intransitive verbs pattern with the direct objects of transitive verbs as being the preferred locus for introducing new mentions. Subjects of transitives, however, are strongly dispreferred slots for the expression of new information. The use of fine-grained grammatical relations enables us to make rather specific claims about the distribution of mentions. The taxonomy of fine-grained grammatical relations is given below in Figure 1.

[1] The feature [Lex] was sufficient for the decision tree tools to learn idiosyncratic uses of gendered pronouns.

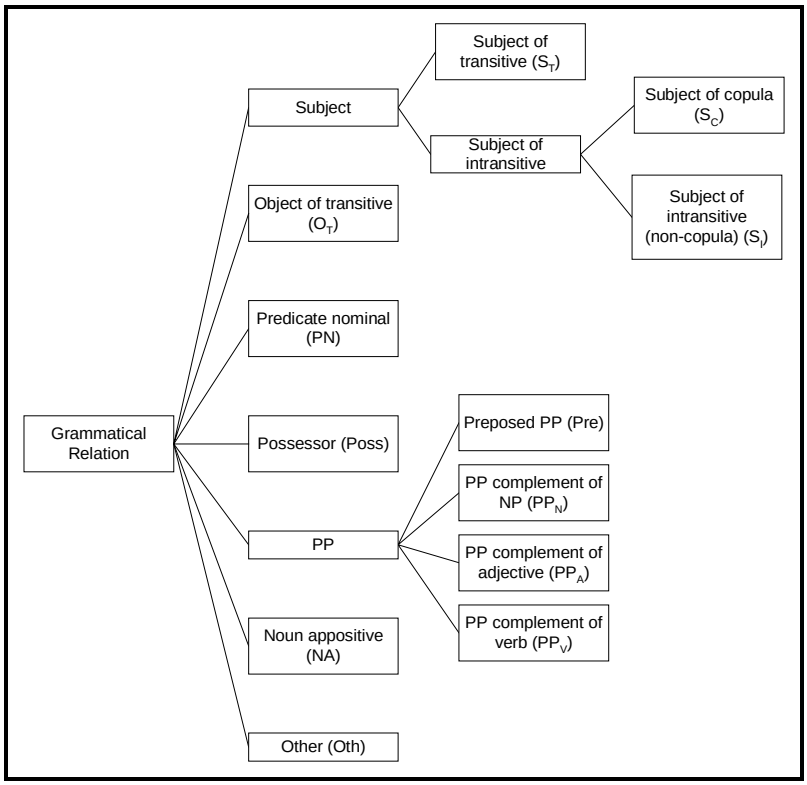

Figure 1 The taxonomy of grammatical relations

4. Decision trees

For a set of annotated examples, we used decision-tree tools to construct the conditional probability of a specific grammatical relation, given other features in the domain.[2] The decision trees are constructed using a Bayesian learning approach that identifies tree structures with high posterior probability (Chickering et al. 1997). In particular, a candidate tree structure (S) is evaluated against data (D) using Bayes' rule as follows:

$$p(S|D) = \text{constant} \cdot p(D|S) \cdot p(S)$$

For simplicity, we specify a prior distribution over tree structures using a single parameter kappa (k). Assuming that N(S) probabilities are needed to parameterize a tree with structure S, we use:

$$p(S) = c \cdot k^{N(S)}$$

where $0 < k \leq 1$, and c is a constant such that p(S) sums to one. Note that smaller values of kappa cause simpler structures to be favored. As kappa grows closer to one (k = 1 corresponds to a uniform prior over all possible tree structures), the learned decision trees become more elaborate. Decision trees were built for $k \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.99, 0.999\}$.

Having selected a decision tree, we use the posterior means of the parameters to specify a probability distribution over the grammatical relations. To avoid overfitting, nodes containing fewer than fifty examples were not split during the learning process. In building decision trees, 70% of the data was used for training and 30% for held-out evaluation.

The decision trees constructed can be rather complex, making them difficult to present visually. Figure 2 gives a simpler decision tree that predicts the grammatical relation of a mention for Encarta at

[2] Comparison experiments were also done with Support Vector Machines (Platt 2000, Vapnik 1998) using a variety of kernel functions. The results obtained were indistinguishable from those reported here.

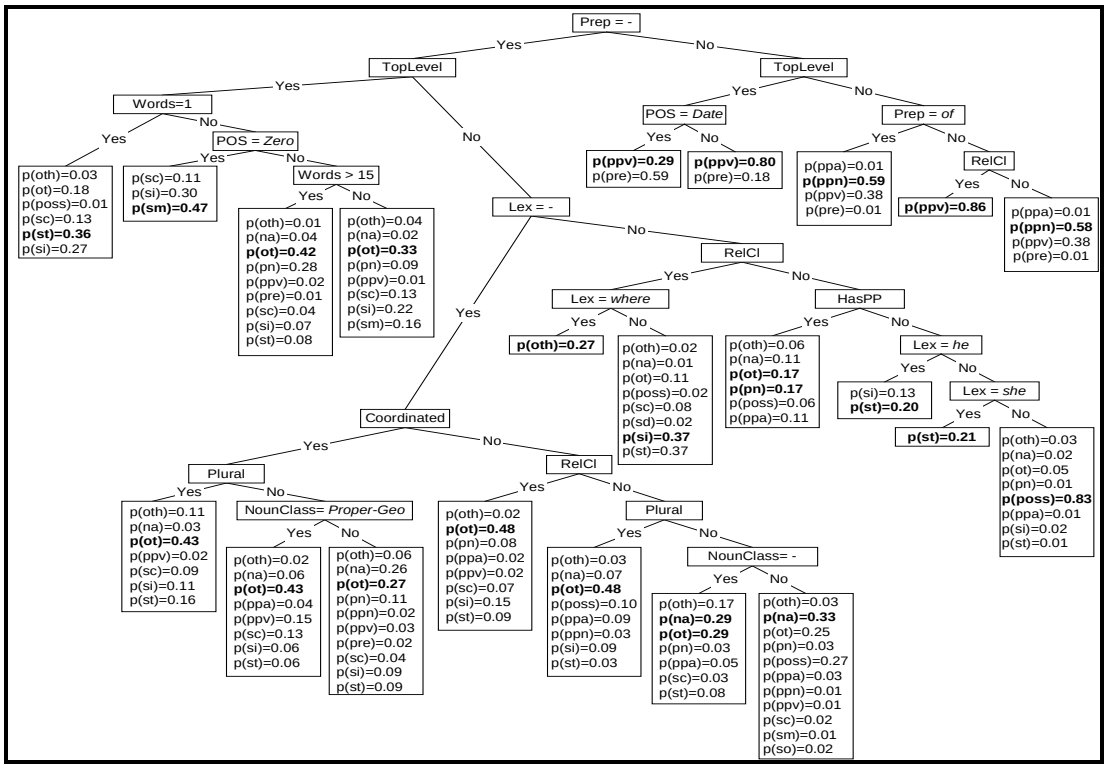

Figure 2 Decision tree for Encarta, at k=0.7

k=0.7. The tree was constructed using a subset of the morphological and syntactic features: [Coordinated], [HasPP], [Lex], [NounClass], [Plural], [POS], [Prep], [RelCl], [TopLevel], [Words]. Grammatical relations with only a residual probability are omitted for the sake of clarity. The top-ranked grammatical relation at each leaf node appears in bold type. Selecting the top-ranked grammatical relation at each node results in a correct decision 58.82% of the time in the held-out test data. By way of comparison, the best decision tree for Encarta computed using all morphological and syntactic features yields 66.05% accuracy at k = 0.999.

The distributional facts about the pronoun "he" represented in Figure 2 illustrate the utility of the fine-grained taxonomy of grammatical relations. The pronoun "he" in embedded NPs ([Prep] = "-", [TopLevel] = No) and when not in a relative clause ([RelCl] = No) favors S_T and S_I. Other grammatical relations have only residual probabilities. The use of the traditional notion of subject would fail to capture the fact that, in this syntactic context, the pronoun "he" tends not to occur as S_C, the subject of a copula.

5. Evaluating decision trees

Decision trees were constructed and evaluated for each corpus. We were particularly interested in the accuracy of models built using only observable features. If accurate modeling were to require more abstract discourse features such as [InformationStatus], a feature that is not directly observable, then a machine-learning approach to modeling the distribution of mentions would not be computationally feasible. Also of interest was the generality of the models.

5.1 Using Observable Features Only

Decision trees were built for Encarta and the Wall Street Journal using all features except the non-observable discourse feature [InformationStatus]. The best accuracy when evaluated against held-out test data and selecting the top-ranked grammatical relation at each leaf node was 66.05% for Encarta at k=0.999 and 65.18% for Wall Street Journal at k=0.99. Previous studies in Preferred Argument Structure (Corston 1996, Du Bois 1987) have

Table 1 Accuracy using only morphological and syntactic features

	Grammatical relations in training data		Accuracy in held-out test data (decision tree accuracy in parentheses)	
Corpus	Top-ranked	Top two	Using top-ranked	Using top two
Encarta	PP_N	PP_N, PP_V	20.88% (66.05%)	41.37% (81.92%)
WSJ	O_T	O_T, PP_N	19.91% (66.16%)	35.56% (80.70%)

established pairings of fine-grained grammatical relations with respect to abstract discourse factors. New mentions in discourse, for example, tend to be introduced as the subjects of intransitive verbs or as direct objects, and are extremely unlikely to occur as the subjects of transitive verbs. Some languages even give the same morphological and syntactic treatment to subjects of intransitives and direct objects, marking them (so called "absolutive" case marking) in opposition to subjects of transitives (so called "ergative" marking). Human referents, on the other hand, tend to occur as the subjects of transitive verbs and as the subjects of intransitive verbs, rather than as objects. Such discourse tendencies perhaps motivate the use of one set of pronouns (the so called "nominative" pronouns {"he", "she", "we", "I", "they"}) in a language like English for subjects and a different set of pronouns for objects (the so called "accusative" set {"him", "her", "us", "me", "them"}). Thus, we can see that distributional facts about mentions in discourse sometimes cross-cut the morphological and syntactic encoding strategies of a language. With a fine-grained set of grammatical relations, we can allow the decision trees to discover such groupings of relations, rather than attempting to specify the groupings in advance.

We evaluated the accuracy of the decision trees by counting as a correct decision a grammatical relation that matched the top-ranked grammatical relation for a leaf node or the second ranked grammatical relation for that leaf node. With this evaluation criterion, the accuracy for Encarta is 81.92% at k=0.999 and for Wall Street Journal, 80.70% at k=0.9.

It is clearly naïve to assume a baseline for comparison in which all grammatical relations have an equal probability of occurrence, i.e. 1/12 or 0.083. Rather, in Table 1 we compare the accuracy to that obtained by predicting the most frequent grammatical relations observed in the training data. The decision trees perform substantially above this baseline. The top two grammatical relations in the two corpora do not form a natural class. In the Wall Street Journal texts, for example, the top two grammatical relations are O_T (object of transitive verb) and PP_N (prepositional phrase complement of a NP). It is difficult to see how mentions in these two grammatical relations might be related. Objects of transitive verbs, for example, are typically entities affected by the action of the verb. Prepositional phrase complements of NPs, however, are prototypically used to express attributes of the NP, e.g. "the man with the red hat". The grammatical relations paired by taking the top two predictions at each leaf node in the decision trees constructed for the Wall Street Journal and Encarta, however, frequently correspond to classes that have been previously observed in the literature on Preferred Argument Structure. The groupings {O_T, S_I}, {O_T, S_C} and {S_I, S_T}, for example, occur on multiple leaf nodes in the decision trees for both corpora.

5.2 Using All Features

Decision trees were built for Encarta and the Wall Street Journal using all features including the discourse feature [InformationStatus]. As it turned out, the feature [InformationStatus] was not selected during the automatic construction of the decision tree for the Wall Street Journal. The performance of the

decision trees on held-out test data from the Wall Street Journal therefore remained the same as that given in section 5.1. For Encarta, the addition of [InformationStatus] yielded only a modest improvement in accuracy. Selecting the top-ranked grammatical relation rose from 66.05% at k=0.999 to 67.32% at k = 0.999. Applying a paired t-test, this is statistically significant at the 0.01 level. Selecting the top two grammatical relations caused accuracy to rise from 81.92% at k=0.999 to 82.23% at k=0.999, not a statistically significant improvement.

The fact that the discourse feature [InformationStatus] does not make a marked impact on accuracy is not surprising. The information status of an NP is an important factor in determining elements of form, such as the decision to use a pronoun versus a lexical NP, or the degree of elaboration (e.g. by means of adjectives, post-modifying PPs and relative clauses). Those elements of form can be viewed as proxies for the feature [InformationStatus]. Pronouns and definite NPs, for example, typically refer to given entities, and therefore are compatible with the grammatical relation S_T. Similarly, long indefinite lexical NPs are likely to be new mentions.

In a separate set of experiments conducted on the same data, we built decision trees to predict the information status of the referent of a noun phrase using the other linguistic features (grammatical relation, clausal status, definiteness and so on.) Zero anaphors were excluded, yielding 4,996 noun phrases for Encarta and 4,758 noun phrases for the Wall Street Journal. The accuracy of the decision trees was 80.45% for Encarta and 78.36% for the Wall Street Journal. To exclude the strong associations between personal pronouns and information status, we also built decision trees for only the lexical noun phrases in the two corpora, a total of 4,542 noun phrases for Encarta and 4,153 noun phrases for the Wall Street Journal. The accuracy of the decision trees was 78.14% for Encarta and 77.45% for the Wall Street Journal. The feature [InformationStatus] can thus be seen to be highly inferrable given the other features used.

5.3 Domain-specificity of the Decision Trees

The decision trees built for the Encarta and Wall Street Journal corpora differ considerably, as is to be expected for such distinct genres. To measure the specificity of the decision trees, we built models using all the data for one corpus and evaluated on all the data in the other corpus, using all features except [InformationStatus]. Table 2 gives the baseline figures for this cross-domain evaluation, selecting the most frequent grammatical relations in the training data. The peak accuracy from the decision trees is given in parentheses for comparison. The decision trees perform well above the baseline.

Table 3 provides a comparison of the accuracy of decision trees applied across domains compared to those constructed and evaluated within a given domain. The extremely specialized sublanguage of Encarta does not generalize well to the Wall Street Journal. In particular, when selecting the top-ranked grammatical relation, the most severe evaluation of the accuracy of the decision trees, training on Encarta and evaluating on the Wall Street Journal results in a drop in accuracy of 7.54% compared to the Wall Street Journal within-corpus model. By way of contrast, decision trees built from the Wall Street Journal data do generalize well to Encarta, even yielding a modest 0.41% improvement in accuracy over the model built for Encarta. Since the Encarta data contains more mentions (5,252 mentions) than the Wall Street Journal data (4,937 mentions), this effect is not simply due to differences in the size of the training set.

Table 2 Cross-domain evaluation of the decision trees

	Grammatical relations in training data		Accuracy in held-out test data (decision tree accuracy in parentheses)	
Train-Test	Top-ranked	Top two	Using top-ranked	Using top two
WSJ-Encarta	O_T	OT, PP_N	15.90% (66.32%)	36.58% (79.51%)
Encarta-WSJ	PP_N	PP_N, PP_V	15.98% (61.17%)	31.90% (77.64%)

Table 3 Comparison of cross-domain accuracy to within-domain accuracy

Top-ranked	
Train on Encarta, evaluate on WSJ	61.17%
Train on WSJ, evaluate on WSJ	66.16%
Relative difference in accuracy	-7.54%
Train on WSJ, evaluate on Encarta	66.32%
Train on Encarta, evaluate on Encarta	66.05%
Relative difference in accuracy	+0.41%
Top two	
Train on Encarta, evaluate on WSJ	77.64%
Train on WSJ, evaluate on WSJ	80.70%
Relative difference in accuracy	-3.74%
Train on WSJ, evaluate on Encarta	79.51%
Train on Encarta, evaluate on Encarta	81.92%
Relative difference in accuracy	-2.94%

5.4 Combining the Data

Combining the Wall Street Journal and Encarta data into one dataset and using 70% of the data for training and 30% for testing yielded mixed results. Selecting the top-ranked grammatical relation for the combined data yielded 66.01% at k=0.99, compared to the Encarta-specific accuracy of 66.05% and the Wall Street Journal-specific peak accuracy of 66.16%. Selecting the top two grammatical relations, the peak accuracy for the combined data was 81.39% at k=0.99, a result approximately midway between the corpus-specific results obtained in section 5.1, namely 81.92% for Encarta and 80.70% for Wall Street Journal.

The Wall Street Journal corpus contains a diverse range of articles, including op-ed pieces, mundane financial reporting, and world news. The addition of the relatively homogeneous Encarta articles appears to result in models that are even more robust than those constructed solely on the basis of the Wall Street Journal data. The addition of the heterogeneous Wall Street Journal articles, however, dilutes the focus of the model constructed for Encarta. This perhaps explains the fact that the peak accuracy of the combined model lies above that for the Wall Street Journal but below that of Encarta.

6. Conclusion

Natural language generation is typically done under one of two scenarios. In the first scenario, language is generated *ex nihilo*: a planning component formulates propositions on the basis of a database query, a system event, or some other non-linguistic stimulus. Under such a scenario, the discourse status of referents is known, since the planning component has selected the discourse entities to be expressed. More abstract discourse features like [InformationStatus] can therefore be used to guide the linguistic encoding decisions.

In the second, more typical scenario, natural language generation involves reformulating existing text, e.g. for summarization or machine translation. In this scenario, analysis of the linguistic stimulus will most likely have resulted in only a partial understanding of the source text. Coreference relations (e.g. between a pronoun and its antecedent)

may not be fully resolved, discourse relations may be unspecified, and the information status of mentions is unlikely to have been determined. As was shown in section 5.2, the accuracy of the decision trees constructed without the feature [InformationStatus] is comparable to the accuracy that results from using this feature, since superficial elements of the linguistic form of a mention are motivated by the information status of the mention.

The decision trees that were constructed to model the distribution of NPs in real texts can be used to guide the generation of natural language, especially to guide the selection among alternative grammatical ways of expressing the same propositional content. Sentences in which mentions occur in positions that are unlikely given a set of linguistic features should be avoided.

One interesting problem remains for future research: why do writers occasionally place mentions in statistically unlikely positions? One possibility is that writers do so for stylistic variation. Another intriguing possibility is that statistically unusual occurrences reflect pragmatic markedness, i.e. that writers place NPs in certain positions in order to signal discourse information. Fox (1987), for example, demonstrates that lexical NPs may be used for previously mentioned discourse entities where a pronoun might be expected instead if there is an episode boundary in the discourse. For example, a protagonist in a novel may be reintroduced by name at the beginning of a chapter. In future research we propose to examine the mentions that occur in places not predicted by the models. It may be that this approach to modeling the distribution of mentions, essentially a machine-learning approach that seeks to mine an abstract property of texts, will provide useful insights into issues of discourse structure.

References

Chickering, D. M., D. Heckerman, and C. Meek, 1997, "A Bayesian approach to learning Bayesian networks with local structure," In Geiger, D. and P. Punadlik Shenoy (eds.), *Uncertainty in Artificial Intelligence: Proceedings of the Thirteenth Conference*, 80-89.

Corston, S. H., 1996, *Ergativity in Roviana, Solomon Islands*, Pacific Linguistics, Series B-113, Australia National University Press: Canberra.

Du Bois, J. W., 1987, "The discourse basis of ergativity," *Language* 63:805-855.

Fox, B.A., 1987, Discourse structure and anaphora, Cambridge Studies in Linguistics 48, Cambridge University Press, Cambridge.

Heidorn, G., 1999, "Intelligent writing assistance," To appear in Dale, R., H. Moisl and H. Somers (eds.), A Handbook of Natural Language Processing Techniques, Marcel Dekker.

Platt, J., N. Cristianini, J. Shawe-Taylor, 2000, "Large margin DAGs for multiclass classification," In Advances in Neural Information Processing Systems 12, MIT Press.

Vapnik, V., 1998, Statistical Learning Theory, Wiley-Interscience, New York.