Preprint of: Ou, S., Khoo, C., Goh, D.H., & Heng, H.Y. (2004). Automatic discourse parsing of sociology dissertation abstracts as sentence categorization. In I.C. McIlwaine (Ed.), *Knowledge Organization and the Global Information Society: Proceedings of the Eighth International ISKO Conference* (pp. 345-350). Wurzburg, Germany: Ergon Verlag.

Automatic Discourse Parsing of Sociology Dissertation Abstracts as Sentence Categorization

Authors:

Shiyan Ou (email: pg00096125@ntu.edu.sg)
Christopher S.G. Khoo (email: assgkhoo@ntu.edu.sg)
Dion H. Goh (email: ashlgoh@ntu.edu.sg)
Hui-Ying Heng (email: ps7610453J@ntu.edu.sg)

Authors' address:

Division of Information Studies
School of Communication & Information
Nanyang Technological University
31 Nanyang Link
Singapore 637718
Tel: (65) 67906564   Fax: (65) 67927526

Shiyan Ou, Christopher S.G. Khoo, Dion H. Goh, Hui-Ying Heng
Division of Information Studies
School of Communication and Information
Nanyang Technological University, Singapore

Automatic Discourse Parsing of Sociology Dissertation Abstracts as Sentence Categorization

Abstract: We investigated an approach to automatic discourse parsing of sociology dissertation abstracts as a sentence categorization task. Decision tree induction was used for the automatic categorization. Three models were developed. Model 1 made use of word tokens found in the sentences. Model 2 made use of both word tokens and sentence position in the abstract. In addition to the attributes used in Model 2, Model 3 also considered information regarding the presence of indicator words in surrounding sentences. Model 3 obtained the highest accuracy rate of 74.5 % when applied to a test sample, compared to 71.6% for Model 2 and 60.8% for Model 1. The results indicated that information about sentence position can substantially increase the accuracy of categorization, and indicator words in earlier sentences (before the sentence being processed) also contribute to the categorization accuracy.

1. Introduction

This paper reports our initial effort to develop an automatic method for parsing the discourse structure of sociology dissertation abstracts. This study is part of a broader study to develop a method for multi-document summarization. Accurate discourse parsing will make it easier to perform automatic multi-document summarization of dissertation abstracts.

In a previous study, we determined that the macro-level structure of dissertation abstracts typically has five sections (Khoo, Ou & Goh, 2002). In this study, we treated discourse parsing as a text categorization problem - assigning each sentence in a dissertation abstract to one of the five predefined sections or categories.

Decision tree induction, a machine-learning method, was applied to word tokens found in the abstracts to construct a decision tree model for the categorization purpose. Decision tree induction was selected primarily because decision tree models are easy to interpret and can be converted to rules that can be incorporated in other computer programs. A well-known decision-tree induction program, C5.0 (Quinlan, 1993), was used in this study.

2. Previous Studies

Discourse structure usually has the form of a tree structure, resulting from the recursive embedding and sequencing of discourse units (Kurohashi & Nagao, 1994). According to Mann & Thompson (1988), a discourse unit has an independent functional integrity, and can be a clause in a sentence, a single sentence, a text segment containing several sentences, or a paragraph. To understand a text, it is important to parse the discourse structure, and identify how discourse units are combined and what kind of relations they have. Discourse parsing algorithms using various kinds of lexical and syntactic clues have been developed by researchers, such as Kurohashi & Nagao (1994), Marcu (1997), and Le & Abeysinghe (2003).

There has been an increasing interest in applying machine learning to discourse parsing, including supervised and unsupervised methods. Nomoto & Matsumoto (1998) used C4.5

decision tree induction program to develop a model for parsing the discourse structure of news articles. Marcu (1999) used C4.5 to develop a rhetorical parser to identify the discourse units of unrestricted texts. Supervised learning gives good results but requires a large training corpus and manual assignment of predefined category labels to the training dataset.

This study applies decision tree induction to categorize sentences, as a method for parsing the macro-level discourse structure of dissertation abstracts in sociology.

3. Data Preparation

A sample of 300 abstracts was selected systematically from the set of PhD dissertation abstracts indexed under Sociology in the Dissertation Abstracts International Database, published in 2001. The sample abstracts were partitioned into a training set of 200 abstracts used to construct the classifier, and a test set of 100 abstracts to evaluate the accuracy of the constructed classifier. All the abstracts were segmented into sentences using a computer program, and the sentences in the abstracts were manually assigned to one of the five predefined categories: *background*, *problem statements*, *research methods*, *research results*, and *concluding remarks.* To simply the classification problem, each sentence was assigned to only one category, though actually some sentences could arguably be assigned to multiple categories or no category at all. Some of the abstracts were found to be unstructured and difficult to code into the five categories. There were 29 such abstracts in the training set and 16 in the test set. The unstructured abstracts were deleted from the training set.

To prepare data for the experiments, the sentences were tokenized and words were stemmed using the Conexor parser (Pasi Japanainen & Timo Jarvinen, 1997). A small stoplist comprising prepositions, articles and auxiliary verbs were used. The word frequency was calculated for each unique word, and only words above a specific threshold value were retained in the study. Different threshold values were explored. Each sentence was converted into a vector of term weights. Binary weighting was used, i.e. a value of “1” was assigned to a word if it occurred in the sentence, “0” otherwise. The dataset was formatted as a table with sentences as rows and words as columns.

4. Experiments

A well-known decision-tree induction program, C5.0 (Quinlan, 1993), was used in the study. 10-fold cross-validation was used to estimate the accuracy of the decision tree built using the training sample, while reserving the test sample to evaluate the final model.

Preliminary experiments (using 10-fold cross-validation) were carried out to determine the appropriate parameters to use in the model-building. The number of minimum records per branch was set at 5 to avoid overtraining. To make it easier to incorporate the output model into other computer programs later, we specified the resulting model to be a ruleset. Boosting was found to contribute little to the accuracy of discourse parsing, and was not employed in the final experiments.

In this study, three models were investigated:

- Model 1 made use of word tokens found in the sentence.
- Model 2 made use of both word tokens and sentence position in the abstract. The position of the sentence was normalized by dividing the sentence number by the total number of sentences in the abstract.
- Model 3 took into consideration indicator words found in other sentences before and after the sentence being categorized, in addition to the attributes used in Model 2.

4.1 Model 1 –- words present in the sentence

Model 1 used high frequency words present in the sentences as the attributes to build the decision tree. The threshold value for the word frequency determines the number of the attributes used in the model. We tested the estimated accuracy of Model 1 with pruning severity of 90%, 95% and 99% separately using 10-fold cross validation for various threshold values. A higher pruning severity results in a smaller and more concise decision tree with a shorter training time. The results are reported in Table 1.

Table 1. Estimated accuracy of Model 1 for various word frequency threshold values

Word frequency threshold values	Number of words input	Pruning Severity		
		90%	**95%**	**99%**
>5	1463	53.7	53.9	53.9
>10	876	54.4	54.4	53.7
>20	454	56.4	55.6	56.3
>35	242	57.5	**57.9**	56.2
>50	153	56.5	56.4	55.5
>75	75	51.6	51.0	50.7
>100	44	51.1	50.8	50.1
>125	30	50.7	50.7	50.7

**The values are estimated accuracy using 10-fold cross validation.*

The results showed that Model 1 obtained the best estimated accuracy of 57.9%, with word frequency threshold value of 35 and pruning severity of 95%. The high word frequency threshold of 35 indicates that only high frequency words are useful for categorizing the sentences. In fact, only a small number of indicator words were selected by C5.0 to develop the decision tree (e.g. 20 indicator words were used in the best model).

After building the final decision tree for Model 1, we applied it to the test sample of 100 abstracts (including 16 unstructured abstracts). The accuracy rate obtained was 50.04%. When the 16 unstructured abstracts were removed from the test sample, the accuracy rate became 60.84%. This means that if we can do some preprocessing to filter out the unstructured abstracts, the categorization accuracy can improve substantially.

4.1. Model 2 -- sentence position

For Model 2, we investigated whether sentence position is helpful in predicting the category of the sentences. The normalized sentence position was used as an additional attribute to build Model 2. As with Model 1, word frequency threshold of 35 was used. The estimate accuracy rates using 10-fold cross validation for various pruning severity values are given in Table 2.

Table 2. Estimated accuracy of Model 1 and Model 2 for various pruning severity

Word frequency threshold values	Number of words input	Sentence position as an additional attribute	Pruning Severity				
			80%	**85%**	**90%**	**95%**	**99%**
>35	242	No (Model 1)	57.0	57..9	57.5	57.9	56.2
		Yes (Model 2)	66.5	66.4	65.1	66.6	65.1

**The values are estimated accuracy using 10-fold cross validation.*

With sentence position as an additional attribute, the estimated accuracy obtained by Model 2 increased substantially. Clearly, sentence position is important in identifying which category or section a sentence belongs to. A common sequence for the five categories in a dissertation abstract is: *background* -> *problem statements* -> *research methods* -> *research results* -> *concluding remarks*.

Pruning severity has not much effect on the accuracy of both Model 1 and Model 2. We selected 95% as the appropriate pruning severity because the training time is shorter, the size of the decision tree is smaller, and it avoids overtraining.

Using 95% pruning severity and 242 high frequency words occurring in more than 35 sentences as well as normalized sentence position as attributes, we constructed the final decision tree classifier for Model 2. Some of rules in the resulting ruleset are shown in Table 3. We applied Model 2 to the test sample of 84 abstracts (not including 16 unstructured abstracts). The accuracy rate obtained was 71.59%, much better than 60.84% for Model 1 (See Table 4).

Table 3. Some of Rules found in Model 2

Rules for Section 1	Rules for Section 2	Rules for Section 3	Rules for Section 4	Rules for Section 5

if N_SENTEN <= 0.444444 then 1 (836, 0.355)
…
	if STUDY = 1 and N_SENTEN <= 0.444444 and PARTICIP = 0 and DATA = 0 and CONDUCT = 0 and PARTICIPATE = 0 and FORM = 0 and ANALYSIS = 0 and SHOW = 0 and COMPLETE = 0 and SCALE = 0 then 2 (172, 0.733)
	…
		if DATA = 1 and TEST = 0 and EXAMINE = 0 and METHOD = 0 and ASSESS = 0 and EXPLORE = 0 then 3 (93, 0.613)
		…
			if REVEAL= 1 and IMPLICAT = 0 then 4 (44, 0.932)
			if SHOW = 1 then 4 (57, 0.842)
			if IMPLICAT = 0 then 4 (2030, 0.41)
				if IMPLICAT = 1 then 5 (33, 0.788)
				if FUTURE = 1 and N_SENTEN > 0.444444 then 5 (36, 0.694)

Table 4. Comparison of sections assigned by Model 1 and Model 2

Section	No. of sentences	Model 1 correctly classified	Model 2 correctly classified
1	173	12 (6.94%)	123 (71.10%)
2	183	98 (53.56%)	102 (55.74%)
3	189	80 (42.33%)	94 (49.74%)
4	468	426 (91.03%)	410 (87.61%)
5	29	16 (55.17%)	17 (58.62%)
Total	1042	634 **(60.84%)**	746 **(71.59%)**

4.2. Model 3 -- indicator words found in surrounding sentences

The dissertation abstract is a continuous discourse with relations between sentences. Surrounding sentences before and after the sentence being processed can help to determine the category of the sentence. For example, if the previous sentence is the first sentence in the *research results* section, then the current sentence is likely to be under *research results* as well. Furthermore, sentences which are easy to classify, because they contain clear indicator words, can be used to help identify the categories of other sentences that do not contain clear indicator words. For example, the *research results* section often begins with a sentence containing clear indicator words, e.g. "*Results showed that …", "The result indicated that …", "The analysis revealed that …", "The study suggested that …", "This study found that …"* . Subsequent sentences will amplify on the results but may not contain a clear indicator word.

To test this assumption, we extracted indicator words from the decision tree of Model 1 and Model 2 (see Table 5). For each sentence, we then measured the distance between the sentence and the nearest sentence (before and after) which contained each indicator word. Table 6 illustrates this. Sentence 13 in document 4 is being processed. The indicator word "*study*" is found in sentence 4 (9 sentences earlier) and sentence 7 (6 sentences earlier), as well as in sentence 14 (1 sentence after).

Table 5. Indicator words found in Model 1 and Model 2

	Model	Number of words	Indicator words
Common words	**Model 1 & 2**	13	complete, conduct, data, dissertation, examine, explore, future, implication, interview, investigate, participate, reveal, test
Unique words	**Model 1**	7	literature, purpose, population, question, qualitative, reform, survey
	Model 2	12	access, age, analysis, form, method, participant, perception, scale, second, show, status, study

Table 6. Indicator words in surrounding sentences

Doc_id	Sentence_id	Neighboring sentence_id	Indicator word	Distance	Location
4	13	4	study	-9	before*
4	13	7	analysis	-6	before
4	13	14	study	1	after*

** "Before" means that the indicator word is in the sentence before the sentence being processed.*
** "After" means that the indicator word is in the sentence after the sentence being processed.*

Then, we used the surrounding indicator words as additional attributes (distance as the attribute values) in 3 ways:

- Sentence position of indicator words *before* the sentence being processed;
- Sentence position of indicator words *after* the sentence being processed;
- Sentence position of indicator words both *before and after* the sentence being processed.

The evaluation results for Model 3 using 84 structured test abstracts are shown in Table 7. Table 7 shows that only indicator words *before* the sentence being processed can contribute to the categorization accuracy (obtaining the best result 74.47%). With indicator words *after* the sentences being processed, the result (68.62%) is even worse than that for Model 2 (71.59%).

Table 7. Test results for Model 3 based on the test sample of 84 structured abstracts

Section	No. of sentences	Model 2 correctly classified	Model 3 correctly classified		
			With all indicator words	Only with before indicator words	Only with after indicator words
1	173	123(71.10%)	140 (80.92%)	138 (79.77%)	117 (67.63%)
2	183	102 (55.74%)	89 (48.63%)	96 (52.46%)	90 (49.18%)
3	189	94 (49.74%)	99 (52.38%)	99 (52.38%)	74 (39.15%)
4	468	410 (87.61%)	426 (91.03%)	426 (91.03%)	418 (89.31%)
5	29	17 (58.62%)	17 (58.62%)	17 (58.62%)	16 (55.17%)
Total	1042	746 **(71.59%)**	771 **(73.99%)**	776 **(74.47%)**	715 **(68.62%)**

5. Conclusion and future work

In this study, we investigated the use of decision tree induction to parse the macro-level discourse structure of sociology dissertation abstracts. We treated discourse parsing as a sentence categorization task. The attributes used in constructing the decision tree models were stemmed words that occurred in more than 35 sentences (out of 3694 sentences in 300 sample abstracts). Sentence position information was found to increase the categorization accuracy rate from 60.8% (Model 1) to 71.6% (Model 2).

We also developed Model 3 that made use of information regarding the presence of 32 indicator words in surrounding sentences. We found that only indicator words *before* the sentence being processed contribute to the categorization accuracy, obtaining the best result of 74.5%.

In future, we plan to carry out more in-depth error analysis to determine whether some inference method can be used to improve the categorization. Other machine-learning methods such as support vector machine (SVM) and Bayesian learning will also be investigated. In addition, the manual categorization of the sample abstracts was done by one person. We plan to have two more codings so that inter-indexer consistency can be calculated, and compared with the performance of the automatic categorization. Finally, we plan to develop a preprocessing program for filtering out the unstructured abstracts to improve the categorization accuracy.

References

Khoo, Christopher, Ou, Shiyan, & Goh, Dion. (2002). A hierarchical framework for multi-document summarization of dissertation abstracts. In *Proceedings of the 5th Conference on Asian Digital Libraries (ICADL-2002).* Singapore. Pp. 99-110.

Kurohashi, Sadao & Nagao, Makoto. (1994). Automatic detection of discourse structure by checking surface information in sentences. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING--94) (vol. 2).* Kyoto, Japan. Pp. 1123-1127.

Le, Huong T. & Abeysinghe, Greetha. (2003). A study to improve the efficiency of a discourse parsing system. In *Proceedings of the 4th International Conference on Intelligent Text Processing and Computational Linguistics (ClCLing-2003).* Mexico City, Mexico. Pp. 356-369.

Mann, W.C. & Thompson, S.A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text,* 8(3), 243-281.

Marcu, D. (1997). The rhetorical parsing, summarization, and generation of natural language texts. PhD Dissertation, Department of Computer Science, University of Toronto.

Marcu, D. (1999). A decision-based approach to rhetorical parsing. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99).* Maryland. Pp.365-372.

Nomoto, Tadashi & Matsumoto, Yuji. (1998). Discourse parsing: a decision tree approach. In *Proceedings of the 6th Workshop on Very Large Corpora (WVLC-98).* Montreal, Quebec, Canada. [http://acl.ldc.upenn.edu/W/W98/W98-1125.pdf]. Accessed 08/25/2003.

Pasi Japanainen and Timo Jarvinen. (1997). A non-projective dependency parser. In *Proceedings of the 5th Conference on Applied Natural Language Processing*. Washington D.C.: Association for Computational Linguistics. Pp. 64-71.

Quinlan, J.R. (1993). *C4.5: programs for machine learning.* San Mateo: Morgan Kaufmann Publishers.