Deletion of Support Verbs During Abstracting

Choy-Kim Chuah

School of Computer Science
Universiti Sains Malaysia
11800 Penang, Malaysia
e-mail : kimc@cs.usm.my

Abstract

For researchers in disambiguation, generation and translation, the machine treatment of idioms and collocations has always been a challenge. However, in a study of the process of content condensation in abstracting, the problem was found not to be as serious, at least for one particular type of collocation. In the most explored of approaches in automatic summarization, sentences are extracted to constitute a summary. To improve on the summary, non-salient words should be deleted. But, what just are these non-salient words? In this paper, we studied one particular type of collocation in the context of abstracting by an author in a scientific domain. Our study found that for verbal complexes such as *cause decreases* and *give shape to*, the verb is more often than not deleted and the noun following it denominalised and used as substitute for the whole complex. As support verbs are semantically empty, this operation does not bring about any change in meaning. The observation is an important finding for automatic summarization. The contexts where support verbs are found are determinable. Support verbs may then included as one of several items on a checklist of linguistic units for condensation.

1 Introduction and Motivation

Since Luhn's (1958) first work, research on automatic summarisation has mainly focussed on the extraction of sentences to constitute a summary (see Edmundson, 1969; Rush *et al*., 1971; Mathis *et al*., 1973; Ono *et al*., 1994; Kupiec *et al*., 1995; Brandow *et al*., 1995; Marcu, 1997; Barzilay & Elhadad, 1997). However, to advance summarisation from sentence selection into sentence condensation, some researchers have in the last 2-3 years begun to identify the linguistic processes involved in content condensation (see Jing & McKeown, 2000; Saggion, 2000; Chuah, 2001). While we know what some of the processes are, we are still unable to reformulate selected content properly. The task of summary writing is a complex one.

In her work, Chuah (2001) identified some linguistic units which are frequently deleted by an author during abstracting[1]. Unlike *substitution* and *aggregation*, the process of *deletion* is less problematic to implement. However, the problem lies in knowing which linguistic units to delete, the context of their deletion and how to recognise them. In this paper, we extend on Chuah's work with an empirical study on one of these units, namely collocations involving support verbs.

1.1 Collocation with Support Verb

In language, some words co-occur only with certain words. In English, *give* co-occurs with *shape*, or *form* to produce the collocations of *give shape to* and *give form to* respectively. While the category of words which may combine with *give* to produce such a collocation

[1] In the context of this paper, the term *abstracting* is used synonymously and interchangeably with *summarisation*. *Abstracts* are for us *summaries* of a special type found in scientific documents.

is not quite as restricted as in idioms, not any word can combine with *give* either. *Give* does not collocate with *contour*: **give contour to*, even though all three words of *shape*, *form* and *contour* are more or less synonymous (for a discussion of collocations see McKeown & Radev, 2000).

Consider the following possible context where the verbal complex just mentioned might be found. The sentence before the arrow was taken from a full text, and that after, its abstract.

(0) Processes of mere exposure, Pavlovian conditioning and social learning **give shape$_N$ to** relationships between liking and other determinants.

→ Processes such as mere exposure, Pavlovian conditioning and social learning **shape$_V$** the relationships between these factors, food liking and eating behavior.

A machine treatment of such collocations would raise problems in analysis: should *shape* be analysed as the direct object of *give*? or should *relationships* be the direct object of *give shape to*?

Grefestette & Teufel (1995) who were working on the automatic identification of support verbs for nominalisation said that support verbs which may be "used with a nominalized predicate structure [are] unpredictable". We are unaware of any known list of verbs which may function as support verbs.

1.2 Motivation

To contribute to current work on condensation, the aim of this study is three-fold: (a) to determine the significance of complexes with a support verb in abstracting, i.e. if such complexes are frequently occurring; (b) to determine empirically the list of collocations involving support verb for natural language processing, and (c) to investigate how author-abstractors handle such collocations during abstracting. The work reported in this paper is preliminary.

In Section 2, we describe the study method. Some statistics on the corpus are also provided. In Section 3, we list the collocations with a support verb that were implicated in abstracting in our study. The paper ends with a discussion and suggestions for future work.

2 Method

2.1 Corpus

Fifty-seven scientific articles from two journals, *Behavioral Ecology and Sociobiology*, and *Oecologia* (Springer-Verlag Publications) were downloaded for the study. All documents were accompanied by author-written abstracts. Some statistics on the corpus are given in Table 1.

Table 1. Statistics on corpus

	full-text (ft)	abstract(ab)	RF
Corpus size	7938 sn; 175,613 wd	534 sn; 11,975 wd	15:1; 15:1
Size of article	62–269 sn; 1,552-6,333wd	5–21 sn; 109-415 wd	7:1–31:1; 7:1–31:1
Av. size of article	139 sn; 3,081 wd	9 sn; 210 wd	15:1; 15:1
Range of sn length	4–129 wd	7–80 wd	

sn = sentence; wd = word; Av. sn length = 22 wds; Reduction factor, RF = No. ft-sn (or wd): No. ab-sn (or wd).

2.2 Identification of selected sentences

Without direct access to how an author abstracts, we have no recourse but to an indirect method. On the basis of verbatim matches, similarity in stem and in meaning, a manual search was made for sentences from full text most likely to have been used in abstracting by an author.

Full text sentences that most resemble given sentences from the accompanying abstract were assumed to have been used in abstracting by the author. A search was then made for collocations involving a support verb in the sentences from both a full text and its abstract. Next, by comparing the selected sentence from full text with its corresponding sentence from abstract, we were able to determine how such a linguistic unit was handled by an author during abstracting.

3 Results

3.1 SUPPORT_VERB + $N_X \rightarrow V_X$

Of the eleven examples of collocations with a support verb (SUPPORT_VERB) found, nine had the support verb deleted during abstracting, and the noun (N_x) denominalised to a verb (V_x)

which then substitutes for the whole verbal complex (see examples (1) to (9)).

To summarise this observation, we write:

$$\text{SUPPORT\_VERB} + N_x \rightarrow V_x \qquad \text{(A)}$$

The common subscript indicates that the noun and the verb share the same stem.

(1) The observation that fluctuations of CO_2 concentration around a plant **lead to a** pronounced **reduction of** oviposition indicates that …

→ On host plants exposed to rapid fluctuations in CO_2 concentration, the frequency of oviposition **was reduced by** a factor of 3.2 compared to the control.

[oec2-97110539][2]

(2) The complex leaf litter habitat of *S. ocreata* may **create an** important physical **constraint** on the effectiveness of vibrational signalling; …

→ vibratory communication is **constrained** by the complex leaf litter habitat of some populations.

[bes1-9638017]

(3) Simple movement, … may **be acquired through a** gradual associative **learning process**, …

→ These patterns may be innate, or they may **be learned** through the bees' early foraging experience.

[bes2-9639293]

(4) These results **provide** strong collective **support for** the basic tenet of the model, that genetic variation in worker task choice influences task flexibility.

→ Our data **support** previous findings that genotypic variation plays an important role in task regulation.

[bes2-9946171]

(5) Surprisingly, however, sensory organs that **are specialised to the detection of** CO_2 find their strongest expression in the almost exclusively herbivorous Lepidoptera.

→ Sensory organs that **detect** CO_2 are common in herbivorous moths and butterflies, but their function has been unclear until now.

[oec2- 97110539]

For about half of the examples involving a support verb in our study, deletion was further followed by substitution with a lexically related word such as its troponym[3] or hypernym[4] (see examples (6) to (9)). We propose the process to have occurred in two steps (see equation A'), and the two verbs, V_x and V_y, to be held by some lexical relation, **R**: $\mathbf{R}(V_x) = V_y$.

$$\text{SUPPORT\_VERB} + N_x \rightarrow V_x \rightarrow V_y \qquad \text{(A')}$$

(6) The presence of fish **caused decreases in** both mating frequency and mating duration,

→ The presence of fish **reduced** both the number of matings per pool (mating frequency), and mean mating durations.

[oec2-97117258]

Troponym(decrease_V) = reduce_V (WN)

(7) **results in a decrease** in photosynthetic surface of …

→ **reduce** the leaf area by …

[oec2-97112209]

Troponym(decrease_V) = reduce_V (WN)

(8) Cannibalism can also **have an impact on** the size structure of populations.

→ Cannibalistic tendencies are well known in spiders and may be a significant factor **influencing** population size.

[bes1-9945349]

Troponym(impact_V) = influence_V (WN)

(9) Several results in this study **argue in favour of** the hypothesis of host heterogeneity.

→ Our results **confirm** this host heterogeneity.

[oec2-98114382]

Hypernym(favour_V) = permit_V (WN)

Hypernym(confirm_V) = permit_V (WN)

[2] The code indicates the source document from which the sentences were taken.

[3] A verb expressing a specific manner elaboraton of anoher verb. X is a troponym of Y if to X is to Y in some manner (from WordNet v. 1.6)

[4] The generic term used to designate a whole class of specific instances. Y is a hypernym of X if X is a (kind of) Y (from WordNet v. 1.6).

3.2 $V_X \rightarrow$ SUPPORT_VERB + N_X

For two examples in our study, the process was the converse of the condensation process just described:

$V_x \rightarrow$ SUPPORT_VERB + N_x (B)

The single simple verb V_x (e.g. *demonstrates*) was replaced by a SUPPORT_VERB+NOUN complex (i.e. *represents the demonstration of*). The noun N_x (i.e. *demonstration*) is a nominalized form of the original verb V_x (i.e. *demonstrates*) (see examples (10) and (11)).

(10) This study **demonstrates that** a geographic structure of host plant specificity can occur in a polyphagous grasshopper.
→ This study **represents the** first **demonstration of** a geographic structure of host plant specificity in a polyphagous grasshopper.
[oec2-99120437]

(11) *C. lineolata* ants **prefer** *A. flavokansiensis* eggs enclosed with excrement from females that feed on bundleflower more than eggs enclosed with excrement from females that feed on other hosts.
→ *C. lineolata* **exhibited a** strong **preference for** eggs from female *A. flavokansiensis* that fed exclusively on bundleflower compared to eggs from females that fed exclusively on honey locust.
[oec1-98115434]

4 Discussion

4.1 How support verbs are handled during abstracting

In her study, Chuah (2001:119-141) identified some linguistic units that were commonly deleted. The degree of deletion for any given unit was observed to be dependent on the unit in consideration.

For example, while connectives are almost always deleted, the modifier *significant* was retained 2-3 times more often than it was deleted (*ibid*.:133). Illocution markers[5], e.g. *We found that*, also exhibited a strong tendency to be deleted. Adverbials, lexical unit *both*, quantifiers, determiners, and hypernyms, on the other hand, exhibited equal tendencies for insertion as well as deletion.

In our study, SUPPORT_VERB+NOUN complexes were about four times more likely to be rendered concise than to be rendered "wordy" by an author himself during abstracting (see section 3). This observation is consistent with abstracting, where we want to rid a text of non-salient words.

Note that in such collocates, the meaning is essentially unaffected even after the verb is deleted, and the deverbalised noun used as replacement for the whole collocate. This observation may be attributed to the fact that support verbs are semantically empty.

Because the deletion of support verbs does not bring about significant changes in meaning, we propose that support verbs be deleted without exception during summarisation. In this way, a collocation, and hence, the text in which it is found, is condensed.

For about half of the examples, the substitute verb does not share stem with the object of the support verb, but is a word that is in some lexical relation with it, often its troponym. Consider the verb *reduced* and the noun *decreases* in example (6), where $reduce_V$ is a troponym of $decrease_V$.

While we are unable to say for sure why substitution was carried out, we propose that the process be effected only if the word has been used in the same or previous sentence to avoid repetition.

If natural language processors have access to a list of such collocates with support verbs, the problem of incorrect object attachment may be circumvented by pre-treating this special construct.

While we consider the observation to be the result of two processes, i.e. deletion followed by substitution, both processes may equally be seen as a single process of substitution of a complex with a single simple verb.

4.2 Empirical list

While our study did not turn up many examples, and we are unable to make any definitive statement about the semantics of the verbs

[5] which "make explicit [the] speech or discourse act [being] perform[ed] at certain points" (Vande Kopple's (1985:83-85).

involved, the verbs appear to concern the interpretation of observations made by the author e.g. *lead to a reduction*, *argue in favour of*, and do not seem to be domain-restricted.

The empirical list of verbal complexes involving support verbs are as follows.

lead to a reduction → reduce
cause a decrease → reduce
result in a decrease → reduce
provide support for → support
represents a demonstration of → demonstrate
acquire through learning → learn
argue in favour of → confirm
have an impact on → influence
exhibited a preference for → prefer
create a constrain → constrain
specialised to the detection of → detect

While the support verb that may be used with a nominalized form is unpredictable (Grefestette & Teufel, 1995), the local context where a support verb may occur is determinable.

From the list of examples obtained from our study, we arrived at seven patterns where support verbs may be found (see Table 2). The patterns are given in terms of grammatical categories. These contextual formats where a support verb may occur is determinable, and we believe the set of patterns to be finite. From our study, the most common pattern is that of V + ART + N + PREP.

Table 2. Formats of support verb complexes

FORMAT[+]	FREQUENCY
V + ART + N	2
V + ART + N + PREP	3
V + PREP + ART + N	2
V + PREP + ART + N + PREP	1
V + PREP + N	1
V + PREP + N + PREP	1
V + N + PREP	1

[+]V = verb; ART =article; N = noun; PREP = preposition

The identification of such a list of patterns containing a support verb is important especially at this time when we are unable to produce proper summaries. It helps to be able to condense small bits of linguistic units to improve on the quality of extracted summaries. Also, the examples themselves provide the context where a support verb may be found, and the support verb to be deleted.

4.3 Future work

In our study, the determination of sentences in full text that were used in abstracting was manually carried out (see sub-section 2.2). Because of the tedium of the task, it was difficult to increase the corpus size.

For a comprehensive study with larger and different corpora, we propose that the task be automated, even if in a semi-automatic way. Once sentences implicated in abstracting are identified, we propose having another program to detect verbal complexes with a support verb using the patterns in Table 2.

With this, future work may be carried out on other corpus types to determine if such collocations are as consistently and similarly reduced, besides adding to our modest list of examples of such verbal complexes.

Acknowledgement

We would like to thank Springer-Verlag Publications for permission to reprint bits of text taken from articles used in our study.

References

Barzilay, R. & Elhadad, M. 1997. Using Lexical Chains for Text Summarization. *Proc. of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pp 10-17, Universidad Nacional de Educación a Distancia, Madrid, Spain, July 11.

Brandow, R., Mintze, K. & Rau, L.F. 1995. Automatic Condensation of Electronic Publications by Sentence Selection. *Information Processing and Management*, 31(5):675-685.

Chuah, C.-K. 2001. *Linguistic Processes for Content Condensation in Abstracting Scientific Texts.* PhD. Thesis. Département de linguistique et de traduction, Faculté des arts et des sciences, Université de Montréal, Montréal, Canada, August 2001.

Edmundson, H.P. 1969. New Methods in Automatic Abstracting. *Journal of the ACM*, 16(2):264-285.

Grefestette, G. & Teufel, S. 1995. Corpus-based Method for Automatic Identification of Support Verbs for Nominalizations Proceedings of EACL 1995, Dublin. Available from the Computation and Language e-print archive as cmp-lg/9503010. (http:www.cogsci.ed.ac.uk/~simone/support.ps.gz)

Jing, H. & McKeown, K.R. 2000. Cut and Paste Based Text Summarization. *Proc. of the*

1st Meeting of the North American Chapter of the Assoc. for Computational Linguistics, pp 178-85, Seattle, Washington, USA. April 29-May 4, 2000.

Vande Kopple, W. 1985. Some Exploratory Discourse on Metadiscourse. *College Composition and Communication*, 36(1):82-93.

Kupiec, J., Pedersen, J., & Chen, F. 1995. A Trainable Document Summarizer. *Proc. of the 18th Annual International ACM SIGIR Conference on Research and Development on Information Retrieval*, pp 68-73. Seattle, Washington USA. July 1995.

Luhn, H.P. 1968. The Automatic Creation of Literature Abstracts. In C.K. Schultz (ed.) H.P. Luhn: Pioneer of Information Science, Selected Works. New York: Spartan Books. pp 118-125. [originally published in 1958 in the *IBM Journal of Research and Development*, 2(2):159-165.]

Marcu, D. 1997. From Discourse Structures to Text Summaries. *Proc. of the ACL'97/EACL'97 Workshop on Intelligent Scalable Text Summarization*, pp 82-88, Universidad Nacional de Educación a Distancia, Madrid, Spain. July 11.

Mathis, B., Rush, J.E., & Young, C.E. 1973. Improvement of Automatic Abstracts by the Use of Structural Analysis. *Journal of the American Society for Information Science*, pp 101-109.

Kathleen R. McKeown & Dragomir R. Radev. 2000. Collocations. In Robert Dale, Hermann Moisl & Harold Somers (eds.) *A Handbook of Natural Language Processing*. Marcel Dekker.

Ono, K., Sumita, K. & Miike, S. 1994. Abstract Generation Based on Rhetorical Structure Extraction. *Proc. of the International Conference on Computational Linguistics*. (COLING-94), pp 344-348. Japan.

Rush, J.E., Salvador, R. & Zamora, A. 1971. The Abstracting and Indexing. II. Production of Indicative Abstracts by Application of Contextual Inference and Syntactic Coherence Criteria. *Journal of the American Society for Information Science*, pp 260-274.

Saggion, H. 2000. *Génération automatique de résumés par analyse sélective*. PhD thesis. Université de Montréal.