Just What May be Deleted or Compressed in Abstracting?

Choy-Kim CHUAH[1]

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

Résumé – Abstract

Les résumés constitués de phrases extraites d'un texte contiennent souvent des mots inutiles; il est possible de les éliminer ou d'en réduire le nombre. Par un étude comparative des phrases du documents et des phrases correspondantes dans le résumé, cet article présente un inventaire partiel des unités qui sont souvent éliminées ou réduites en nombre. Les metadiscours, les textes entre parenthèses, les unités redondantes (emphases, répétitions), les appositions, modifieurs et relatives ne sont pas à place dans un résumé.

Abstracts constituted from extracted sentences contain unneeded information that may be deleted, or compressed into simpler units. By comparing full text sentences used in abstracting with correspond-ing sentences in abstract, the study found such units to include metadiscourse phrases, parenthetical texts, redundant units inserted for emphasis, or are repetitions. Apposed texts and units such as modifiers and relative clauses which provide details and precision in the full text, but are out of place in an abstract, are also deleted.

Mots-clés - Keywords

abstracting, deletion, compression, metadiscourse, redundancy

1 Introduction

Abstracts constituted from extracted sentences (Kupiec *et al*., 1995; Marcu, 1997; Barzilay & Elhadad, 1997) are not only disjointed, but also contain unneeded texts that may be deleted or compressed. To improve on cohesion and coherence of an extracted abstract, Jing (2000:311-312) removes as many "extraneous phrases" as possible based on "context weight" calculated for each word. While Jing's results are optimistic, what would be desirable is if superfluous linguistic units can be semi-automatically pruned off using a checklist, and reduction effected on compressible units. So, just what are some of these units that may be deleted, or compressed?

While Saggion (2000:58-59) did identify four types of deletion, our study will seek to extend on this operation, and identify the function of the units involved. We will also discuss compression during which inessential lexical units are removed from a linguistic unit. The basic meaning is however left unaltered in the indispensable units. The types of unit and

[1] A PhD student at Département de linguistique et de traduction, Université de Montréal.

context under which linguistic units may be condensed are important considerations for abstracting systems. However, our study is preliminary and a proper investigation is needed.

2 Method

Fifty-seven articles from two on-line journals on biology were used for the study. All sentences in full text (ft-) and abstract (ab-) were given a code indicating its location in the document. For example, a sentence with location code [R-2-1] is the first sentence in the second paragraph of the Results section (I = Introduction; M = Method; D = discussion; A = abstract). On the basis of verbatim matches, similarity in stem and meaning, a manual search was made for ft-sentences that were probable sources of information in abstracting. Examples are presented in the following format: <ft-sentence> → <ab-sentence>; REL(ft-LU) = ab-LU. The lexico-semantic relation (REL) between a deleted and undeleted word is looked up in WordNet (WN), and given in the notation of the Explanatory-Combinatorial Dictionary (Melcuk *et al.*, 1995; Melcuk, 1996).

3 Results and Discussion

3.1 Deletion

3.1.1 Metadiscourse

Metadiscourse is "Writing about writing, whatever does not refer to the subject matter being addressed" (Williams, 1981 cited by Vande Kopple, 1985). Vande Kopple (*ibid.*:83) explains that an author usually writes at two levels. At one level, propositional content on a subject is supplied, and at another, metadiscourse which does not contribute to the content, but helps a reader "interpret … [the] material" is added. The implication here is that if superfluous metadiscourse can be reliably deleted, then what is left is the propositional content.

Illocution markers which "make explicit [the] speech or discourse act [performed] at certain points" (*ibid.*:83-85), and connectives are two types of metadiscourse commonly ~~deleted~~ during abstracting (ϕ = omission). Closely associated with illocution markers are first person pronouns. Forty percent (22/55[2]) of documents exclude an author's overt presence from the abstract, by using the agentless passive and personifying inanimate nouns (see (2)) which is typical of scientific writing. The group of verbs involved include *study* and *observe*. As much as metadiscourse does not contribute to propositional content, not all types of metadiscourse are deleted to the same degree during abstracting. Unlike illocution markers, connectives in the source sentences are almost always deleted. The information selected for inclusion in an abstract is then reformulated and re-connected with "new" connectives.

(1) ~~In the present study~~$_x$, ~~we~~$_y$ investigated the effects of larval shields of *Cassida* spp. that feed
upon tansy towards the ant *Myrmica rubra*, … [I-3-1]
→ ϕ_x ϕ_y The effects of these abdominal shields towards *M. rubra* were studied in three
cassidine species, … [A-1-4; oec2-98118166]

(2) ~~Based on~~$_x$ our observations, ~~we~~$_y$ suggest that this is due to interference competition … [I-5-4]
→ ϕ_x Our observations ϕ_y suggest that … may be due to interference competition …
[A-1-10; oec2-97109313]

Because connectives are identifiable, and *we* and *I* are the two main overt indicators of author, both metadiscourse types may be deleted from extracted sentences without any need

[2] Two documents did not contain first person pronouns, *I* and *we*.

for computation of salience, which is a desirable feature for automatic summarizers. However, a study into the degree to which illocution markers may be deleted is necessary. According to Meyers (1992:297), metadiscourse phrases such as *We found that*, are stereotypical means of making "strong, distinctive, but polite claims".

3.1.2 Precision and Details

Quantifier, determiner, apposed text

The restricted length of an abstract has little place for precision and details. During abstracting, determiners, quantifiers, including nouns of measure, e.g. *density*, and numeric expressions which provide details may be omitted. Deletion is however not consistent. In (3), measure noun *density* was deleted, but not in (11). Linguistic units most reliably deleted are parenthetical and apposed texts (see (4)). A deleted meaning may be re-expressed elsewhere (see subscripted units in (3)).

(3) there was a highly significant$_x$ decline in ~~the overall density of~~$_y$ the plants. [R-3-2]
→ The ϕ_x decline ϕ_y of the weed has been most evident$_x$ … [A-1-7; oec1-98114343]

(4) the host ranges of two ~~species of~~$_x$ chrysomelid beetles~~, *Ophraella notulata* and *O.slobodkini*~~$_y$
that are specialized on different species in the Asteraceae. [I-6-1]
→ in the host specialization of two ϕ_x chrysomelid beetles ϕ_y that are monophagous on
different species of Asteraceae. [A-1-1; oec2- 97112081]
Hypernym(*Ophraella*) = beetle (domain knowledge);

NOUN$_1$-of-NOUN$_2$ (N$_1$-of-N$_2$) construction

In NOUN$_1$-of-NOUN$_2$ constructions, a deleted unit can be the lexical noun before or after *of*. In (5), while the first noun is dispensable, deletion of the second noun leads to ungrammaticality: **influenced only one aspect* ϕ. In (6), where either may be omitted, the more detailed is deleted, it suffices as is often the case, to keep the noun that expresses the "gist".

(5) influenced ~~only one aspect of~~ male activity or microhabitat use; [R-3-5]
→ influenced ϕ water strider behavior; [A-1-9; oec2-97117258]

(6) A recent study ~~of the life history of this annual species~~ revealed an unusually extended
reproductive period, which results in a very wide and possibly bimodal size distribution of the
coexisting juvenile instars. [I-6-2]
→ Preliminary field observations ϕ indicated an extended reproductive period, which results
in a very wide size distribution of juvenile instars. [A-1-3; bes1-9945349]

While deletion is not linked to position, neither does it appear to be word-dependent except for *species* where it is deleted in two clear situations for the corpus studied: (a) if the meaning of phrase *species of X*, is the same as *X* itself (see (7)), and (b) if it is clear that the noun in question is a species name, i.e. *R. alternata* (see (8)).

(7) specializes on a few ~~species of~~ moths [I-2-2]
→ attract certain ϕ male moths [A-1-3; oec1-97112572]

(8) Many ~~species of~~ the tephritid genus *Rhagoletis* are very common. [I-2-2]
→ *Rhagoletis alternata* is a common ϕ tephritid fly [A-1-1; oec1-98115154]

Prepositional modifiers and adverbials which provide details are often deleted, but may also be appended to add disparate information about the experiment.

3.1.3 Implicit knowledge

A modifier containing information that is discoverable from the head noun or context, may be deleted. While the social nature of spider mites in (9) may be found from domain, in (10), the semantic component of *immature* in *young* is implicit in *larvae*.

(9) In a ~~subsocial~~ spider mite, *Schizotetranychus miscanthi* Saito, … [I-2-1]
→ the ϕ spider mite, *Schizotetranychus miscanthi* … [A-1-1; bes1-9946025]

(10) The presence of ~~young~~ larvae also affects the proportion of foragers collecting pollen:..[I-2-5]
→ The decision to collect pollen by honey bee foragers depends on the number of ϕ larvae (brood), … [A-1-2; bes2-9844193]
Syn(young) = immature (WN) {**Syn** = synonym}; larva = immature free-living form (WN)

In scientific reportage, it is implicit that observations must be *significant* to be reported, and that assessments when made are *relative* to some referent. Modifiers *significant* and *relative* were retained (2-3 times) more often than they are deleted. While implicit, the modifiers lend credence to the results. As much as findings on the deletion of evaluative modifiers are not conclusive, subjective evaluative modifiers such as *readily* and *dramatic*, are almost always deleted.

3.1.4 Explicitness

Hypernym

A common noun, scientific name, or a technical name may be made more explicit by postposing it with its hypernym which contains redundant information. A hypernym was not found to be deleted, when it postposes a technical name, e.g. *chrysomelid* in (4). However, when it postposes a scientific name (see (11)), it is deleted 44% (8/18) of the time. While it is clear to novice readers that scientific names, which are always in italics or underscored, are proper names, it is not clear what technical names may refer to, and this hampers comprehension.

(11) there was a … decline in the density of the mature *S. punicea* ~~plants~~ … [R-2-2]
→ There has been a … decline in the density of mature *S. punicea* ϕ …
[A-1-6; oec1-98114343]
Hypernym(*S. punicea*) = plant (domain knowledge); **Hypernym**(plant) = life form (WN)

Hypernyms postposed to lexically ambiguous common nouns are not deleted, except in the case of partial repetitions. Where absent, an appropriate one may even be inserted. A common domain-related hypernym for an ACTIVITY noun is *behavior*.

While WordNet appears to be an adequate resource for finding relations between words, even those involving knowledge in biology, a problem that confronts abstracting by sentence extraction is in deciding which among the possible hypernyms to use in a semi-automatic abstracting environment. As some words may have their own domain-related hypernym, a special thesaurus supplemented with domain knowledge will be invaluable.

Contrary to technical nouns which usually have but one restricted meaning, common nouns may be polysemic or lexically ambiguous. To make explicit its meaning, the hypernym cooccurring with a common noun is retained, and where absent, an appropriate one may even be inserted. The non-deletion and insertion of hypernyms is reflective of the readership at which an abstract is directed, and of author's effort to make explicit unfamiliar knowledge to novice readers.

Emphatic *both*

Content selected for inclusion in an abstract may be presumed to be important, and hence, do not require further emphasis. However, to draw a attention to important points, especially if the material is expected to be unfamiliar, emphatics may equally be retained. As with lexical words, emphatic redundant function words, e.g. *both*, too may be deleted (see (12)), or retained (see (13)).

(12) ~~Both~~ undertakers and guards were less likely to **engage in**$_x$ behavior **typical of**$_y$ young bees **than**$_z$ food storers and wax workers. [D-3-2]
→ ϕ Guards and undertakers were less likely to **perform**$_x$ behavior **normally associated with**$_y$ young bees **compared to**$_z$ food storers and wax workers. [A-1-6; bes2-9741151]

3.2 Compression

An alternative way to condense is to compress complex linguistic units (with multiple lexical units) into simpler units (see **underscored text in bold**).

3.2.1 $V_{support} + N \rightarrow \phi + V_0(N)$

In the first group of compressible units, the verb in a VERB+NOUN collocation is semantically empty. The whole verbal complex ($V_{support}$ + N) may be compressed by deleting $V_{support}$, and replacing the noun, N, with its verbal derivation, V_0(N), e.g. *to cause decrease*$_N$ → *to decrease*$_V$. The meaning of the complex is essentially unchanged. Substitution of the derived verb with a synonym usually optionally follows (see (13)).

(13) The presence of fish **caused decreases** in both mating frequency and mating duration, ..[R-6-3]
→ The presence of fish **ϕ reduced** both the number of matings … and mean mating durations. [A-1-15; oec2-97117258]

3.2.2 Catenative + VERB$_{non\text{-}finite}$ → VERB

The second group of compressible units involve a catenative. A catenative is "a lexical verb which governs the non-finite form of another lexical verb" (Crystal, 1997). During abstracting, a catenative which is not semantically empty, may in few and restricted cases be replaced by the non-finite verb if its deletion does not bring about a change in meaning. Note that it is not always possible to know linguistically if the change in meaning is marginal. While *X was allowed to hatch* may be condensed to *X hatched*, we may or may not condense *X tend to move* to *X move*. Only the author-researcher knows if a catenative may be deleted.

While both groups of compressible units just discussed are uncommon in the study corpus (only one example per two documents), it merits an investigation as it involves determinate situations, and may be pertinent in other corpus type. While catenatives are often associated with phasal verbs, e.g. to *start/continue/stop to VERB*, only one example was found in the study.

3.2.3 Derivation, Compound Noun Formation, and Prepositional Phrase Formation

Besides verbal phrases, other groups of words may also be variously compressed by a mix of processes including derivation, e.g. nominalization and adjectivalization, and compound noun formation (see (15)). In (16), note the verb deleted from the VERB+PREPOSITION complex.

(15) environmental conditions that constrain$_V$ vibratory communication [D-1-4]
→ environmental constraints$_N$ [A-1-9; bes1-9638017]

(16) Plants either of whose parents originated from the Bayshore location … than … [D-1-3]
→ Plants with parents from one of three locations … [A-1-5; oec1-99120268]

4 Concluding Remarks

Linguistic units commonly deleted include illocution markers containing first person pronouns, connectives, parenthetical texts, apposed texts and repetitions. While deletion of such linguistic units may be a first step in condensation, multiple deletions of such units alone can significantly abridge a text without critical loss in core content. However, the situation is usually much more complicated than implied in (18).

(18) although ~~high~~ temperatures ~~clearly~~ had a ~~suppressive~~ effect ~~on foraging~~. [R-3-3]
→ although ϕ temperature ϕ had some ϕ effect ϕ. [A-1-10; oec2-98117420]

While an investigation into factors critical for reliable identification of dispensable units is invaluable to condensation, other studies should include the extent to which an author's presence may be deleted, and deletion in N_1-of-N_2 constructions according to noun type, in general and in other domains. Pending such long-term studies, short-term projects can find units which may be deleted or compressed. A list of nouns for which *behavior* is its hypernym is also useful.

Références

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.

Crystal, D. (1997) *A Dictionary of Linguistics and Phonetics*, 4th edn. Oxford: Blackwell.

Jing, H. (2000) Sentence Reduction for Automatic Text Summarization. Proc. of *the 6th Applied Natural Language Processing Conference*, pp. 310-315, Seattle, Washington, USA. April 29-May 4, 2000.

Vande Kopple, W.J. (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.

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.

Melcuk, I. (1996) Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In L. Wanner (ed.), *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdan/Philadelphia: Benjamins, pp. 37-102.

Melcuk, I., Clas, A. & Polguère, A. (1995) *Introduction à la lexicologie explicative et combinatoire*. Bruxelles: Duculot.

Meyers, G. (1992) Strategic Vagueness in Academic Writing. In Ventola, E. & Mauranen, A. (eds.) *Academic Writing: Intercultural and Textual Issues*, Amsterdam/ Phildelphia: John Benjamins Publishing Company, pp 3-17.

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