Aggregation by Conflation of Quasi-Synonymous Units in Author Abstracting

Choy-Kim CHUAH

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

Département de linguistique et de traduction
Université de Montréal
Montréal (Québec) H3C 3J7, Canada
choy.kim.chuah@umontreal.ca

Résumé – Abstract

La plupart du temps, les études qui portent sur l'aggrégation des phrases en génération de texte, se focalisent sur l'utilisation des connecteurs pour relier les phrases courtes et inventées. Mais, les connecteurs limitent le nombre des unités qu'il est possible de combiner à la fois. Comment condenser l'information en peu d'unités, sans utiliser trop de connecteurs ? Cette étude porte sur des documents ayant trait à la biologie et discute de l'agrégation des phrases par les auteurs quand ils résument. Cet article présente aussi quelques préalables et difficultés pour un système de résumé automatique. Beaucoup de phrases sont aggrégées sans signe explicite, ni connecteur, ni ponctuation.

In text generation, studies on aggregation often focus on the use of connectives to combine short made-up sentences. But connectives restrict the number of units that may be combined at any one time. So, how does information get condensed into fewer units without excessive use of connectives? From a comparison of document and abstract, this reconnaissance study reports on some preferred patterns in aggregation when authors write abstracts for journal articles on biology. The paper also discusses some prerequisites and difficulties anticipated for abstracting systems. More sentences were aggregated without than with the use of an explicit sign, such as a connective or a (semi-)colon.

Mots-clés - Keywords

abstracting, sentence, aggregation, synonymy, conflation

1 Introduction and Motivation

Research on automatic abstracting has mainly stopped at *content selection*[1] (see Kupiec *et al.*, 1995; Marcu, 1997; Barzilay & Elhadad, 1997). While there are notable contributions on

[1] Summarization proper may be divided into two stages of content selection and content condensation.

aggregation from text generation (see Dalianis, 1999; Shaw, 1998; Dalianis & Hovy, 1993), the work is of little immediate benefit to summarization by sentence extraction. The simple made-up sentences aggregated (see Dalianis, 1999:386) do not reflect the complex state of affairs in documents such as scientific and technical journal articles with urgent need for abstracts[2]. If studies on summarization are to benefit real applications, then research must reflect real contexts.

In the aforementioned studies, sentences are almost always aggregated with the use of an explicit sign, a connective or a (semi-)colon. But explicit signs restrict the number of units that may be combined at any one time. So, the question asked is how does information get condensed and without excessive use of explicit signs as observed in scientific abstracts?

1.1 This study

This motivated the present study to investigate how authors combine segments of texts, ultimately sentences, when abstracting scientific journal articles. This reconnaissance study on abstracting reports on some preferred patterns in aggregation, and prerequisites and difficulties that an abstracting system might face. The paper does not propose which pattern or connective to use, both of which depends in part on what and how one wants to communicate selected information. While Section 2 describes the study method, Section 3 gives some data on the distribution of number of full text (ft-) sentences used to write a sentence in an abstract (ab-) and common patterns in abstracting. The paper ends with a discussion and suggestions for future work.

1.2 Aggregation: A Definition

Dalianis & Hovy (1993:90), who worked on "removal of redundancy" during generation, say that Mann & Moore (1980) were the first to use aggregation, although Paice (1981) was reported (in Paice, 1990:175) to have coined the term to mean "the idea of adding adjacent sentences". To refer to this sub-process in the context of abstracting, the term with the underlying meaning of 'combining sentences' will be retained, but with the stipulations of "removal of redundancy", and "adjacent" removed. Not only is deletion not necessarily implicated[3], but sentences aggregated need not be adjacent. For a survey of definitions for aggregation, see Reape & Mellish (1999).

2 Method

Fifty-seven articles from two journals, *Behavioral Ecology and Sociobiology* (bes) and *Oecologia* (oec) (Springer-Verlag Publications) were downloaded for the study. All articles have the basic sections of Abstract (A), Introduction (I), Method (M), Results (R) and Discussion (D). For identification purposes, all sentences in full text and abstract were given a code which indicates 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. On the basis of verbatim matches, similarity in stem and in meaning, a manual search was made for ft-sentences that were probable sources of information for abstract[4] (henceforth, *selected ft-*

---

[2] Which is used here to refer to the special kind of *summary* in scientific and technical documents.

[3] Consider the trivial example where no linguistic unit is deleted: *The elephant is big* + *The mouse is small* → *The elephant is big, but the mouse is small*.

[4] The present author is a trained entomologist.

sentences). For some statistics on the corpus, see Table A1 in the Appendix. Examples provided are in the following format:

(0) <ft-sentence> [location code]
→ <ab-sentence > [location code; journal-year_volume_page]
REL(ft-LU) = ab-LU

3 Results

3.1 Ft-sentences in aggregation

3.1.1 Distribution

About 2/3 of ab-sentences in the study corpus were aggregated from multiple sentences: 37% from two ft-sentences, while another 27% were constituted from three or more sentences, which is an indication of the importance of multiple sentence aggregation in abstracting. For about one-third of corpus, about half of the ft-sentences selected for abstracting came from the Introduction section, none from the Method section (see Chuah, 2001:59).

Table 1. Distribution of ft-sentences to construct an ab-sentence

Sub-corpus (no. ab-sentence)	No. ft-sentence (%)			
	1	2	3	≥ 4
bes1 (120)	43 (35.83)	48 (41.67)	20 (16.67)	9 (7.50)
bes2 (120)	43 (35.83)	48 (40.00)	14 (11.67)	15 (12.50)
oec1 (136)	43 (31.62)	45 (33.09)	30 (22.06)	16 (11.77)
oec2 (158)	60 (37.97)	56 (35.44)	25 (24.05)	14 (8.86)
Corpus (534[+])	189 (35.39)	197 (36.89)	89 (16.67)	54 (10.11)

[+] Five ab-sentences did not have matches.

3.1.2 Source

As this is a preliminary study, and because of the complexity of the problem, we only looked at the simplest case of two-sentence aggregation to determine the source of information. Are they from the same section? Different sections? Most sentences aggregated were from the same section with Introduction as the highest contributor, and Method, the lowest. When from different sections, the sentences were likely to be from the Results and Discussion sections.

Table 2. Distribution of selected ft-sentences in two-ft-one-ab-sentence construction

Section	Section			
	Introduction	Method	Results	Discussion
Introduction	57[+](28.9)			
Method	12 (6.1)	10 (5.1)		
Results	5 (2.5)	9 (4.6)	31 (15.7)	
Discussion	11 (5.6)	5 (2.5)	27 (13.7)	30 (15.2)

[+] No. of sentences (percentage)

Eighteen percent of ab-sentences has its source in sentences that were immediately adjacent. The implication for aggregation is that adjacent sentences are more likely to be on

the same topic than sentences from different paragraphs/sections, and the anaphor probably refers to an element mentioned in the preceding sentence.

3.2 Categorization of aggregation

Reape & Mellish (1999:23-25) proposed a four-category typology for aggregation. Conceptual aggregation was distinguished from semantic and lexical aggregations. While the latter two presumptively involve linguistic knowledge, the examples given do not appear to be far different from that of conceptual aggregation which implicates world/domain. However, on the basis of whether an explicit sign was used or not, we propose three categories of aggregation. If the explicit sign is a connective or (semi-)colon, then CONNECTIVE or (SEMI-)COLON respectively, and if no sign was used, then CONFLATION. In the last category of CONFLATION, the basis of aggregation is knowledge, linguistic or world/domain. Refer to Table 3 to see how our proposed categorization compares with that by Reape & Mellish (1999). Each of these categories, C1-C3, is discussed below.

Table 3: Categorization proposed by present study vs. Typology of aggregations surveyed by Reape & Mellish (1999)

Proposed category	Reape & Mellish's typology
By conflation	Conceptual aggregation, e.g. peacock + hummingbird → bird
	Semantic aggregation, e.g. J is C's sister + C is J's brother → C and J are brother and sister
	Lexical aggregation, e.g. Monday + …Friday → weekdays
	Referential aggregation, e.g. John is here + Jane is here → They are here
With connective	Discourse aggregation, e.g. (see Reape & Mellish, 1999:23)
	Syntactic aggregation, e.g. John is here + Jane is here → John and Jane are here
With (semi-)colon	-

3.2.1 By conflation

Seventy-five percent of two-sentence aggregations were the result of *conflation* (see **text in bold**). Two semantically equivalent text units may be conflated by: (a) splicing and joining, or (b) merging them. Units are merged on the basis of semantic similarity. Often one sentence (Sx) is used as the main sentence.

C1a: $[X_1Y]_{Sx} + [X_2Z]_{Sy} \rightarrow [X_2Y]_S$	$'X_1' \cong 'X_2'$[5];

In (1), text unit ***small, early-instar bolas spiders*** was spliced off one sentence and joined to text unit *of both sexes attract moth flies in the genus Psychoda* in main sentence [I-3-6].

(1) **Small, early-instar bolas spiders** ~~do not capture moths~~. [I-3-1]
juvenile bolas spiders of both sexes attract adult male flies in the genus *Psychoda*. [I-3-6]
→ **Small, early-instar bolas spiders** of both sexes attract moth flies in the genus *Psychoda*, … [A-1-5; oec1-97112572]

[5] X, Y, Z are units of text, and 'X' = meaning of X.

C1b: $[X_1Y]_{Sx} + [X_2Z]_{Sy} \rightarrow [XY]_S$	'X_1' + 'X_2' ≅ 'X';

In (2), sentences are aggregated when semantically equivalent text units were merged, before being optionally followed by other condensation sub-processes, such as deletion (~~deleted text~~) and substitution.

recent study + field studies → recent field studies → preliminary field observations

In both cases, units are aggregated without any explicit use of a connective, or a (semi-)colon.

(2) ~~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]
Field studies ~~have suggested that size difference might be important in wolf spider cannibalism~~. [D-1-4]
→ **Preliminary field observations** indicated an extended reproductive period, which results in a very wide size distribution of juvenile instars. [A-1-3; bes1-9945349]

Aggregations, however, are rarely as direct as (1) and (2). In (3), anaphor resolution is required: ***species*** is the lexical anaphor for ***ants and spiders***.

(3) **Ants and spiders** are ~~among the most ubiquitous and diverse predators in terrestrial ecosystems~~. [I-1-1]
Many species ~~share the same trophic level and~~ can potentially compete with and prey upon each other. [I-1-2]
→ **Spiders and ants** are potential competitors and mutual predators. [A-1-1; oec2-97109313]

In (4), experimental knowledge is first required to know that text unit ***CO_2 sensitivity*** is a metonym for text unit ***sensory organs that are specialised to the detection of CO_2***, before a unit was selected. The selected unit was transformed finally to ***sensory organs that detect CO_2*** in the abstract. Note the simultaneous occurrence of other condensation processes, namely substitution with a less technical term: *moths* and *butterflies* → *Lepidoptera*, and compression into fewer words: *functional role* → *function*.

(4) ~~Surprisingly, however~~, **sensory organs that are specialised to the detection of CO_2** find their strongest expression in ~~the almost exclusively~~ herbivorous Lepidoptera. [I-1-7]
~~This suggests that~~ **CO_2 sensitivity** is important throughout that order, but the functional role has remained unclear. [I-1-8]
→ **Sensory organs that detect CO_2** are common in herbivorous moths and butterflies, but their function has been unclear until now. [A-1-1; oec2- 97110539]

3.2.2 With a Connective

Leech & Svartvik (1975:158) listed: coordination, subordination, and adverbial link, as three ways to aggregate clauses. Depending on whether equal, or unequal weight is to be given to the units, the appropriate conjunction, or adverbial is then used.

Aggregation by Coordination

Selected clauses from complex sentences (Sc) are commonly aggregated with a coordinate conjunction, e.g. *and*, *but*, *or*, to form another complex sentence The selected clauses need not share a common unit.

C2a: $[S_1]_{Sc} + [S_2]_{Sc} \rightarrow [S_1$ connective $S_2]_{Sc}$

(5) ~~Facultative slavemakers are able to forage, nurse their brood and construct their nest like free-living ants, and hence~~ colonies without slaves are common. [I-1-3]
Formica subnuda is a facultative slave-making ant, ~~and belongs to the *F. sanguinea* group~~. [I-2-1]
→ *Formica subnuda* is a facultative slave-making ant**, and** colonies without slaves are often found. [A-1-1; bes2-9638145]

As sentences studied are highly complex with multiple sentences, it is possible that the units aggregated are from the same sentence.

(6) ~~The four … stimuli derived from this video had different degrees of asymmetry, and were created to address different aspects of~~ asymmetry manipulation: ~~(1) removed: one tuft was removed~~, representing the most extreme level of FA or RA; ~~(2) reduced: one tuft was reduced in height such that the overall area was decreased by 25%~~, representing a mid-point within the range of natural FA variation; ~~(3) enlarged: …; (4) balanced: …~~ [M-6-3]
→ Asymmetry treatments represented values within the range of natural FA variation **as well as** more extreme values characteristic of regenerative asymmetry. [A-1-9; bes1-9945087]

C2b: $[NP_1VP_1]_{Sc} + [NP_1VP_2]_{Sc} \rightarrow [NP_1\ VP_1\ \text{connective}\ VP_2]_{Sc}$

If coordinated aggregation involves a shared unit, then the redundant unit has to be deleted. As in aggregation by conflation, to combine, the abstractor must first determine the units to be equivalent or synonymous: in (7), *parasitism by eulophids* and *eulophid parasitism* are equivalent. Aggregation was followed by a substitution which requires domain knowledge: generic word *taxa* substitutes for *hybrid and parental plants*.

(7) ~~*Phyllonorycter* survival~~, parasitism by eulophids, ~~and unknown causes of mortality~~ varied significantly among ~~naturally occurring~~ hybrid and parental plants in 1994. [D-1-2]
Eulophid parasitism, ~~rather than unknown mortality~~, appeared to account for the variation in survival among taxa. [D-1-3]
→ Parasitism by eulophid wasps differed significantly among taxa in 1994 **and** appeared to account for the variation in their survival. [A-1-3; oec2-97110360]

In (8), aggregation is complicated by anaphor resolution, and knowing when and what may be deleted. While the fact that the sentences here are consecutive, helps to determine the entity referred to by the anaphor *these*, document knowledge is still required to determine what the noun referred to is. Is it *rules*, or is it *process*?

(8) Simple movement rules, ~~such as the two rules described above~~, may be acquired through a gradual associative learning process, ~~such as the learning mechanisms which lead to the formation of flower-species preferences~~. [I-3-1]
An alternative hypothesis is that these are innate, ~~instinctive~~ processes, ~~and thus should be observable in bees with no previous foraging experience~~. [I-3-2]
→ These patterns may be innate**, or** they may be learned through the bees' early foraging experience. [A-1-2; bes2-9639381]

Aggregation by Subordination

The patterns of aggregation for subordinated and coordinated aggregation differ in the choice of conjunction which depends very much on the communicative intent of the author which a non-author abstractor usually has no direct access.

(9) ~~Combined, these two findings suggest that~~ *S.dumicola* has control over its mean sex ratio but not of its variance. [D-5-4]

~~There are two possibilities that are not mutually exclusive: either~~ the sex ratio biasing mechanism in *S.dumicola* cannot be modified to control the sex ~~of individual offspring~~ or the sex ratio variance ~~is selectively neutral in this system~~. [D-6-2]

→ The sex ratio biasing mechanism in this species, **therefore**, apparently only allows control of the mean sex ratio but not of its variance. [A-1-7; bes1-9946237]

3.2.3 With (Semi-)colon

In aggregations with a semi-colon or colon, the punctuation substitutes for the implicit semantic relation which has been expressly omitted. Aggregations with a (semi-)colon as with other aggregation types, are accompanied by various condensation sub-processes.

Semi-colon

Ehrlich & Murphy (1974:111) say that "When no close relationship exists between two independent clauses, a semicolon can be used to join them".

C3: $[X]_S + [Y]_S \rightarrow [X \text{ (semi-)colon } Y]_S$

While this makes the semicolon a convenient means for combining just about any two clauses, most of the clauses aggregated in the present study are related.

(10) The most abundant prey organisms brought to the nest were Aphidoidea (48.1%), followed by Psocoptera (12.5%), ~~and Lepidoptera larvae (6.0%)~~. [R-9-4]
Only three spiders (two lycosids and one salticid) were brought to the nests. [R-9-5]
→ The majority of prey captured by ants were Aphidoidea (48.1%) and Psocoptera (12.5%)**<semi-colon>** spiders represented only 1.4% of the ants' diet. [A-1-8; oec2-97109313]

In one of four cases, the (semi-)colon is additionally accompanied by a connective to make explicit the semantic relation (see (11)).

(11) ~~*E.snoddyi* males that obtained copulations and unsuccessful males that did not obtain copulations were analyzed to determine if~~ male body size or male balloon size were important criteria for male mating success. [R-4-1]
The empty balloon produced by some species of empidine flies has been hypothesized to be a sexually selected trait. [D-4-1]
→ Both male body size and balloon size are important components in determining male mating success; **<semi-colon>** **however,** the empty balloon does not appear to play a typical role as a sexually selected ornament. [A-1-11; bes1-9945161]

Colon

Colons are used "to set off a series of words, phrases, or clauses from the rest of a sentence, to restate, explain or illustrate a statement immediately before it; … to replace a semicolon for stylistic purposes [to break between clauses]" (Ehrlich & Murphy, 1974:25-27). In the study corpus, there were more examples of aggregation with a colon than with a semi-colon.

(12) ~~In this study, we investigate the role played by~~ a conspicuous male secondary sexual characteristic ~~in the courtship~~ of the wolf spider *Schizocosa ocreata* ~~(Hentz)~~ (Araneae: Lycosidae). [I-3-2]
~~Morphologically, these species can be distinguished only by~~ a male secondary sexual characteristic, a conspicuous tuft of bristles and dark pigmentation on the tibia ~~and patella~~ of the first pair of legs of mature male *S.ocreata*, ~~which is lacking in *S.rovneri* (as well as in the females and juveniles of both species)~~. [I-5-3]
→ Males of the brush-legged wolf spider, *Schizocosa ocreata* (Araneae: Lycosidae), possess a conspicuous male secondary sexual character**<colon>** dark pigmentation and tufts of bristles on the tibiae of their forelegs. [A-1-1; bes1-9638017]

4 Discussion

4.1 Occurrence of aggregation

Aggregation is an important sub-process in condensation. Two-thirds of ab-sentences are the result of combining text units from multiple sentences. Of two-sentence aggregations, three quarters were combined without an explicit sign by conflating semantically equivalent units, while the rest were combined with an explicit sign, a connective or a (semi-)colon (in the ratio of 4:1). This shows the importance of studying aggregation without an explicit sign. Most of the sentences aggregated come from Introduction, which is reflective of the section where important sentences might be found.

4.2 Types of Aggregation

4.2.1 By conflation

While Shaw (1998:139) in his study on text generation noted "coordinate constructions [to be] the most popular aggregation operations, followed by PPs [i.e. prepositional phrases], and then adjectives", three per four ab-sentences in the present study were aggregated by conflation. This is not surprising since aggregating with an explicit sign (connective/(semi-)colon), restricts the number of units that may be combined at any one time. If maximum information is to be condensed into a single sentence, then aggregation by conflation is more effective. To avoid the use of excessive explicit signs to combine sentences, studies in generation need to look into aggregation by conflation.

To conflate sentences, a myriad of processes, condensation and non-condensation, are often implicated. Also, the units to be merged must first be determined to be semantically equivalent. As the units are often equivalent under the guise of synonyms, hypernyms, partial repetitions and metonyms, knowledge ranging from linguistic to experimental to world/domain, is prerequisite. The present research which is preliminary, needs to be followed by studies into the determination of equivalent units and the type of knowledge involved.

4.3 With Connective or (Semi-)Colon

To help a reader process a complex sentence on unfamiliar material, aggregations with connectives which make explicit the semantic relation between units joined, are preferred. In such aggregations, even if a non-author abstractor can decide on the pattern of aggregation, the crux of the problem is which conjunction or adverbial to use such that author's intent is communicated.

The use of a (semi-)colon to aggregate sentences does not mean that there is no semantic relation between the sentences, rather, that "the connection is implicit, and has to be inferred by the reader" (Leech & Svartvik, 1975:162). However, because of the need to be explicit in scientific and technical texts, an adverbial or conjunction is additionally inserted 25% of the time to help a reader process unfamiliar text. The sign to use is also contingent on factors such as style and target reader.

Unlike linguistic units combined in made-up sentences, units actually aggregated are not only different in syntactic class, but are from sentences of differing structure and require experimental knowledge to know that they co-refer, e.g. *tritrophic-level interactions* and *herbivore-parasitoid interaction* (see (16)).

(16) Plant hybridization affects tritrophic-level interactions ~~in this system in the field~~. [D-1-1]
~~However~~, the common garden results ~~strongly~~ suggest that the differences in enemy impact among plants has a genetic basis. [D-6-3]
→ The common garden results show that genetic differences in plants affect the herbivore-parasitoid interaction. [A-1-7; oec2-97110360]

4.4 Problems and Prerequisites

Because sentences in scientific and technical documents are not only long but complex, with an average of about 22 words, a simple aggregation is not possible. To ensure that the output sentence is readable, aggregation is almost always accompanied by various condensation sub-processes, e.g. deletion, to prune off marginal texts. Studies into the interplay of these other processes with aggregation is ultimately necessary. An abstract has also to be formulated in words appropriate to the intended readership.

Besides these problems, an abstracting system is faced with problems related to the prerequisites of abstracting, such as anaphor resolution, determination of the entity referred to in metonymy, and determination of the full form of partial repetitions. Even if the problem of anaphor resolution is alleviated when the sentences are adjacent, and the full form of compound nouns can be determined by a simple concordance of relevant nominal forms, the uncovering of an entity referred to in metonymy which requires experimental or world knowledge, remains problematic. To go beyond, solutions to these problems have first to be found.

4.5 Conclusion and Future Work

As just seen, aggregation in real situations is far different from that treated in hypothetical situations. While one may know how to aggregate, and to detect redundancy, the role of experimental and domain knowledge in conflation is equally urgent. Because conflation is an effective and common means of aggregation, future studies should look into the exploitation of knowledge to this end. Pending long-term measures to understand this condensation sub-process, short-term studies can concentrate on condensing single sentences with the ultimate aim of combining them. We note that the present findings on how an author aggregate sentences are more pertinent to summarization by extraction than by generation.

A study situation proposed for further research is scientific and technical articles, which not only have a high turnover and demand, but are a source of examples for finding patterns/strategies in aggregation. The sentences are highly complex. As aggregation involves other condensation sub-process, parallel studies should be conducted to address problems on the (automatic) deletion, and identification of synonymous units, and the entity referred to in metonymy.

Acknowledgments

I would like to thank all my anonymous reviewers for providing feedback which greatly improved the article. The research reported here was made possible by support from Universiti Sains Malaysia.

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.

Chuah, C. (2001) *Linguistic Processes for Content Condensation in Abstracting Scientific Texts*. PhD thesis. Université de Montréal (to appear).

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

Dalianis, H. (1999) Aggregation in Natural Language Generation. *Computational Intelligence*. 15(4):384-415.

Dalianis, H. & Hovy, E. (1993) Aggregation in Natural Language Generation. In Giovanni Adorni and Michael Zock (eds.) *Trends in Natural Language Generation: An Artificial Intelligence Perspective*. Fourth European Workshop, EWNLG '93, Pisa, Italy, April 1993, pp. 88-105.

Ehrlich, E. & Murphy, D. (1974) *Concise Index to English.* New York: McGraw-Hill Book Co..

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

Leech, G. & Svartvik, J. (1975) *A Communicative Grammar of English*. London: Longman.

Mann, W.C. & Moore, J.A (1980) Computer as Author – Results and Prospects. Research report ISI/RR-79-82, University of Southern California Information Sciences Institute, Marina del Rey.

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.

Paice, C.D. (1981) The Automatic Generation of Literature Abstracts: An Approach Based on the Identification of Self-Indicating Phrases. In R.N. Oddy, Robertson, S.E., van Rijsbergen, C.J. & Williams, P.W. (eds.) *Information Retrieval Research*, pp. 172-191.

Paice, C.D. (1990) Constructing Literature Abstracts by Computer: Techniques and Prospects. *Info. Processing & Management* 26(1):171-186.

Reape, M. & Mellish, C. (1999) Just What is Aggregation Anyway? Proc. of *the 7th European Workshop on Natural Language Generation*, pp. 20-29, Toulouse, France, 13-14 May, 1999.

Shaw, J. (1998) Clause Aggregation Using Linguistic Knowledge. Proc. of *the 9th International Workshop on Natural Language Processing*, pp. 138-147, Niagara-on-Lake, Ontario, Canada, 5-7 Aug.

Appendix

Table A1. Statistics on corpus

	Full-text	Abstract	Reduction factor (RF)
Corpus size	7938 sn[+]; 175,613 wd	534 sn; 11,975 wd	7938:534 = 15:1; 15:1
Size of article	62–269 sn; 1,552-6,333 wd	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 = sentences; wd = words; RF = No. ft-sn (or wd): No. ab-sn (or wd). Av. sn length = 22 wds.