Meta-evaluation of Summaries in a Cross-lingual Environment using Content-based Metrics

Horacio Saggion
Department of Computer Science
University of Sheffield
211 Portobello Street
Sheffield S1 4DP
England, UK
saggion@dcs.shef.ac.uk

Dragomir Radev
School of Information & Department of
Electrical Engineering and Computer Science
University of Michigan
550 E. University, 3080 West Hall
Ann Arbor, MI 48109-1092
radev@si.umich.edu

Simone Teufel
Computer Laboratory
Cambridge University
JJ Thomson Avenue
Cambridge CB3 0FD, UK
Simone.Teufel@cl.cam.ac.uk

Wai Lam
Department of Systems
Engineering & Engineering Management
The Chinese University of Hong Kong
Shatin, Hong Kong
wlam@se.cuhk.edu.hk

Abstract

We describe a framework for the evaluation of summaries in English and Chinese using similarity measures. The framework can be used to evaluate extractive, non-extractive, single and multi-document summarization. We focus on the resources developed that are made available for the research community.

1 Introduction

Evaluation is an essential step of any natural language processing task. In the field of text summarization almost all research is published with an in-house evaluation, which makes it difficult to replicate experiments, to compare results, or to use evaluation data for training purposes. The development of standards of evaluation and sharable resources, such as the Document Understanding Conference (DUC, 2000) among others, is of paramount importance for progress in text summarization.

Evaluations can be intrinsic or extrinsic (Sparck Jones and Galliers, 1995): intrinsic evaluation measures the content of the summary by a comparison with an "ideal" or "target" summary. Extrinsic evaluation measures how helpful summaries are in the completion of a given task, for example in question answering or text categorization.

If intrinsic evaluation is performed by comparing extracted sentences to a set of "correct" extracted sentences, then co-selection is measured by precision, recall and F-score (Firmin and Chrzanowski, 1999). But these measures only consider sentence identity and not sentence content to carry out the comparison, which has the following negative effect: if two extracts consist of different sentences, whereby the sentences convey the same meaning, they are judged as very different by this measure, even though intuitively they would be judged as equivalent. As consequence of the fact that these measures consider only binary decisions (a sentence either is or is not in the extract), they ignore partially correct answers. Also, many researchers have opposed these measures; the generally accepted opinion is that there is no such thing as one ideal summary. Instead, a summary consists of a set of main ideas that should be conveyed (Jones and Paice, 1992)

The most extensive extrinsic evaluation of summarization systems was the TIPSTER SUMMAC evaluation (Mani et al., 1998). In that evaluation, given a generic summary (or a full document), a human assessor had to perform different tasks, eg. relevance decision of a document given a query, or categorization of a document into one out of five categories to which the document is relevant. The evaluation seeks to determine whether the summary is effective in capturing whatever information in the document is needed to correctly categorize the document. SUMMAC was extremely labour-intensive because of the need for assessors who had to read each of the full documents or extracts, which is a clear disadvantage of extrinsic measures of evaluation.

In our research we investigated measures for content evaluation based on the notion of vocabulary overlap. They are developed to palliate the problems with precision and recall. As they are completely automatic, they overcome the problems of task-

based evaluations. These metrics are believed to be quite effective in determining the informativeness of a summary (Mani et al., 2001), and can be used in both extractive and non-extractive summarization, single and multi-document summarization. Recent research has shown how content-based evaluation can be carried out in automatic or semi-automatic fashion (Donaway et al., 2000; Paice and Oakes, 1999).

In this paper, we are interested in meta-evaluation: a comparative evaluation of evaluation measures for summarization. We present a framework for evaluation of the content of automatic summaries, which relies on the availability of target summaries and extracts produced by humans. All the data created for this evaluation is available to the community for research purposes (http://www.clsp.jhu.edu/ws2001/groups/asmd).

2 Experimental Framework

The resources used in this research have been constructed in the context of the 2001 Workshop on Automatic Summarization of Multiple (Multilingual) Documents, a 6-week language engineering workshop at the Center for Language and Speech Processing, Johns Hopkins University. The objectives of the workshop were the integration of cross-lingual information retrieval with single and multi-document summarization and its evaluation.

2.1 Data and Annotation

We use a parallel corpus of English and Chinese (Cantonese) texts which are translations or near translations of each other. The corpus consists of 18,461 document-pairs. The corpus, called the *Hong Kong Newspaper Corpus*, is provided by the Linguistic Data Consortium (LDC). We automatically separated the main title from the main body of text of the news article, inserted sentence and word boundaries (Grover et al., 2000). Semi-automatic corrections of sentence boundaries were made in those sets of documents where human sentence segmentation was available. The English corpus was further annotated with part of speech tags (Mikheev, 2000) and morphologic information, and both Chinese and English text were annotated with named entity tags. Sentence-level alignment was performed based on our reimplementation of Gale and Church's (1991) alignment algorithm. For a complete description of the corpus the reader is referred to (Saggion et al., 2002).

We used 400 documents for our experiments. They were clustered into document sets of 10 documents about one subject ("narcotics rehabilitation", "natural disaster victims aided", "customs staff doing good job", etc.). LDC annotators developed 40 such queries according to our guidelines, then they used an in-house information retrieval engine and human revision, to find the 10 most relevant documents for that query. We provided a manual Chinese translation of each query.

Three LDC judges then assessed each sentence in the 10 relevant documents, and assigned each sentence a score on a scale from 0 to 10, expressing how important this sentence is for the summary. This annotation, which is called "utility judgement", allows us to compile human-generated 'ideal' summaries at different compression rates, which is one gold-standard we use for our different measures of sentence-based agreement, both between the human agreement and between the system and the human annotators. We call this gold standard "human extracts".

The judges also wrote multi-document summaries for each cluster at 50, 100, and 200 words (independently of the size of the documents). As human summary writing by trained professionals is very expensive, it was not possible to provide summaries of all 400 documents by several subjects (and several compression rates). However, our judges found the writing of multi-document summaries to be natural task. They followed the DUC guidelines to do so (DUC, 2000). These texts are a different gold standard we use (only for multi-document summaries); we call them "human summaries".

2.2 Content-based Measures

Content-based similarity measures are functions that take as arguments two text representations and compute a real value in the interval [0..1], the value 1 means that the two texts are closely related while the value 0 means that the two texts are quite different. We have specified and implemented the following measures:

Cosine similarity is computed using the following formula (Salton, 1988):

$$cos(X,Y) = \frac{\sum x_i * y_i}{\sqrt{\sum (x_i)^2} * \sqrt{\sum (y_i)^2}}$$

where X and Y are text representations based on a vector space model. We use two possible weighting schemes for the terms: presence/absence of the term in the text or $tf * idf$ computed using corpus and within text term distribution.

Unit overlap is computed using the following formula:

$$overlap(X,Y) = \frac{\|X \cap Y\|}{\|X\| + \|Y\| - \|X \cap Y\|}$$

where X and Y are text representations based on sets. Here $\|S\|$ is the size of set S.

Longest Common Subsequence is computed using the formula:

$$2*lcs(X,Y) = length(X)+length(Y)-edit_{di}(X,Y)$$

where X and Y are representations based on sequences and where $lcs(X,Y)$ is the length of the longest common subsequence between X and Y, $length(X)$ is the length of the string X, and $edit_{di}(X,Y)$ is the minimum number of deletion and insertions needed to transform X into Y (Crochemore and Rytter, 1994). When comparing two texts, we compute a normalized pairwise lcs between the sentences of the two texts. Unlike cosine and overlap, longest common subsequence is sensitive on how information is sequenced in the text.

As an illustration, consider the following two sentence fragments:

(S1) the terrorist attacked the president.

(S2) the president attacked the terrorist.

the longest common subsequence between S1 and S2 has length 3 (because of the matching subsequences "the...attacked the..."), giving a similarity score of 3/5. Cosine similarity and token overlap consider S1 and S2 as "identical" with score 1. Metrics that consider the linguistic sequence, such as n-gram combinations or longest common subsequence can detect the difference in this case.

For each source document and target length, three different target extracts produced from sentence utility judgement exist. Given an automatic extract S, the three target extracts $Judge\ 0$, $Judge\ 1$, $Judge\ 2$, and a similarity measure M we compute the following numbers:

$$M(S, Judge_i)\ i \in \{0,1,2\}$$

$$Average(M,S) = \frac{\sum_{i\in\{0,1,2\}} M(S,Judge_i)}{3}$$

$$Max(M,S) = Max_{i\in\{0,1,2\}}\{M(S,Judge_i)\}$$

$$Min(M,S) = Min_{i\in\{0,1,2\}}\{M(S,Judge_i)\}$$

Average has been used before for content-based measures (Donaway et al., 2000) while Max and Min have been used only for co-selection (Salton et al., 1994).

Dimension	Values considered
Type of structure	set, vector, and sequence
Complexity	unigrams or bigrams
Form	word or lemma
Parts of speech	nouns, verbs, all parts of speech

Figure 1: Dimensions of text representation considered

2.3 Text Representation

One can compare text units at different levels of analysis: For example one can compare units relying on the number of word or token that two units share, or one can compare the number of lemmas they share. One can use only nouns as the representation, based on the idea that are the nouns that carry the content of the sentence; one might alternatively use main verbs. We experimented with all these parameters and allow our measures to operate at different granurality levels (cf. figure 1).

In the case of texts in Chinese, we don't rely on parts of speech, but we do explore words and Chinese characters as possibilities, because we have developed algorithms to deal with these two text representations (http://www.mandarintools.-com/segmenter.html).

2.4 Summarization Technologies

All summarizers considered in this evaluation are sentence extractors, i.e. they take as input a compression rate (n%) and a document (or cluster of documents) split into sentences, and output an n% extract of the document (or cluster of documents). Sentence extraction is a currently wide-spread, useful technique, but more research in summarization now is moving towards summarization by generation (Jing and McKeown, 2000; Saggion and Lapalme, 2000). Two ways of measuring summary length were explored in our framework: in sentences and in words.

In our experiments we used three summarizers: Websumm (Mani and Bloedorn, 1999) for single and multi-document extracts for English texts; Mead (Radev et al., 2000) for English, Chinese, single and multi-document extracts; and Summarist (Hovy and Lin, 1999) for single-document extracts of Chinese texts.

We consider two baselines in our experimental framework (for both English and Chinese): (i) Lead-based summarizer: n% sentences are picked up from the beginning of the text; and (ii) Random summarizer: n% sentences are picked up at random. Random summaries should give a lower bound for

the performance any system should have, while it is well-known that lead-based summaries perform very well for certain text types.

2.5 Example

We present a complete example of our evaluation measures using document 19980306_007.e ("Number of reported drug abusers dropped in 1997"). Figure 2 shows the summaries produced at sentence compression 10% by the three judges and by one of our summarizers. Table 1 shows the similarity between automatic and human extracts. We also include co-selection metrics for comparison purposes. Note that as extracts agree on how many sentences were extracted, Precision, Recall and F-measure are identical. This example clearly shows that there is no agreement between the summarizer and Judge 2 at the sentence level, nevertheless content-based measures show similarity on content. Also, co-selection measures are identical for Judge 0 and Judge 1, but the two target summaries are rather different and content-based measures are able to capture that difference. In Appendix A, we present Chinese versions of the extracts produced for the document 19980306_006.c (the Chinese version of document 19980306_007.e).

3 Content-based results

In this section we give an overview of the results obtained using content-based metrics, bearing in mind that our objective is not to demonstrate that one particular system is better than other, but to create a useful framework for evaluation. The numbers presented here are based on sentence-level compression, words, and all parts of speech. We present numbers for cosine ($tf * idf$) and longest common subsequence. The results obtained for a subset of target lengths using content-based evaluation can be seen in tables 2 and 3. In all our experiments with cosine ($tf * idf$), the lead-based summarizer obtained results close to the human extracts in most of the target lengths while Mead is ranked in second position. In all our experiments using longest common subsequence, results are inconclusive because no system appears to outperform the others.

The experimental framework for evaluation of the Chinese summaries is based on the novel idea of using the aligned corpus as a source for obtaining a target abstract in Chinese. Given a collection of monolingual summaries, we can use our alignment tables to generate reasonable corresponding cross-lingual summaries and use the collection of these "pseudo manual" chinese summaries in our experiments. This was at all possible because of the accuracy of the alignment program: A preliminary evaluation of our alignment algorithm measured precision and recall at 95.5% and 95.5% respectively.

Method	10%	20%	30%	40%
Summarist	0.44	**0.65**	0.71	0.78
Lead Based	**0.54**	0.63	0.68	0.77
Mead	0.49	**0.65**	**0.74**	**0.82**
Random	0.31	0.50	0.65	0.71

Figure 3: Chinese Summaries. Cosine ($tf * idf$). Average over 10 Clusters. Vector space of Words as Text Representation.

Method	10%	20%	30%	40%
Summarist	0.32	**0.53**	0.57	0.65
Lead Based	**0.42**	0.49	0.54	0.64
Mead	0.35	0.50	**0.60**	**0.70**
Random	0.21	0.35	0.49	0.54

Figure 4: Chinese Summaries. Longest Common Subsequence. Average over 10 Clusters. Chinese Words as Text Representation

The numbers obtained in the evaluation of Chinese summaries for cosine ($tf * idf$) and longest common subsequence can be seen in tables 3 and 4. Both measures identify Mead as the summarizer that produced results closer to the ideal summaries (these results were replicated across measures and text representations).

We have based this evaluation on human extracts produced by LDC assessors (and sentence-alignment in the Chinese case). Nevertheless, other alternatives exist: Content-based similarity measures do not require the target summary to be a subset of sentences from the source document, thus, content evaluation based on similarity measures can be done using human-written summaries.

4 Evaluation of Multi-document Summarizers using Human Summaries

In this evaluation we compare human multi-document extracts with human multi-document summaries. We also compare automatic multi-document summaries produced by Mead with human multi-document summaries. As in the single document evaluation, the results for the human extracts are an average because three different multi-document extracts exists for each cluster. The results for all measures can be seen in tables 4 and 5. Not surprisingly, these results show that human extracts are closer to human summaries than automatic extracts are to human summaries. However, human multi-documemt extracts and automatic multi-document extracts are rather similar ac-

Judge 0 (sentences 2, 3, 13):
The number of drug abusers reported to the Central Registry of Drug Abuse (CRDA) in 1997 totalled 17,555, a drop of 10.8 per cent over the 19,671 reported in 1996, according to CRDA statistics presented to the Action Committee Against Narcotics (ACAN) today (Friday). Speaking at ACAN 's quarterly meeting, the Commissioner for Narcotics, Mrs Clarie Lo, said that young drug abusers reported in 1997 had also dropped by 14.3 per cent compared with 1996. Mrs Lo pointed out that there was a small decrease in the number of female drug abusers in 1997 when 2,216 abusers were reported, compared with 2,429 in 1996.

Judge 1 (sentences 1, 2, 15):
Number of reported drug abusers dropped in 1997. The number of drug abusers reported to the Central Registry of Drug Abuse (CRDA) in 1997 totalled 17,555, a drop of 10.8 per cent over the 19,671 reported in 1996, according to CRDA statistics presented to the Action Committee Against Narcotics (ACAN) today (Friday). Notwithstanding this observation, a thorough study on factors affecting the drug abuse trend in Hong Kong was recently commissioned by ACAN with a view to identifying the underlying factors that affect the size, complexity and characteristics of the drug abuse population in Hong Kong.

Judge 2 (sentences 15, 17, 19):
Notwithstanding this observation, a thorough study on factors affecting the drug abuse trend in Hong Kong was recently commissioned by ACAN with a view to identifying the underlying factors that affect the size, complexity and characteristics of the drug abuse population in Hong Kong. On preventive education, Mrs Lo said that more resources would be devoted to stepping up the beat drugs campaign despite the drop in the number of drug abusers figures. (i) to heighten awareness of the undesirable consequences of abusing drugs, no matter 'hard ' or 'soft ';";

MEAD (sentences 2, 26, 27):
The number of drug abusers reported to the Central Registry of Drug Abuse (CRDA) in 1997 totalled 17,555, a drop of 10.8 per cent over the 19,671 reported in 1996, according to CRDA statistics presented to the Action Committee Against Narcotics (ACAN) today (Friday). ACAN was also informed by a Social Welfare Department 's representative at the meeting that the Subventions and Lotteries Fund Advisory Committee had supported the Government to grant $16.12 million to subvent the services provided by four non-medical voluntary drug treatment and rehabilitation agencies for the 1998/99 financial year. The four agencies are the Barnabas Charitable Service Association, the Christian New Being Fellowship, the Finnish Missionary Service Ling Oi Youth Centre and the Operation Dawn, all of which will also be granted a total subvention of $1.26 million to cover their expenses for the month of March in the 1997/98 financial year.

Figure 2: Target and Mead Extracts

MEAD	Judge 0	Judge 1	Judge 2	Max	Min	Average
Precision = Recall = F-measure	0.33	0.33	0.00	0.33	0.00	0.22
Cosine (0/1)	0.55	0.51	0.23	0.55	0.23	0.43
Cosine (tf*idf)	0.54	0.57	0.21	0.57	0.21	0.44
Unigram	0.37	0.33	0.13	0.37	0.13	0.28
Bigram	0.26	0.26	0.03	0.26	0.03	0.18
LCS	0.47	0.50	0.13	0.50	0.13	0.37

Table 1: Summarizer v. human judges

Method	10%	20%	30%	40%	50%
Lead Based	**0.55**	**0.65**	**0.70**	**0.79**	**0.84**
MEAD	0.46	0.61	**0.70**	0.78	0.83
Random	0.31	0.47	0.60	0.69	0.75
Websumm	0.52	0.60	0.68	0.77	0.82

Table 2: English Summaries. Cosine ($tf * idf$). Average over 10 Clusters. Words and all POS as text representation.

Method	10%	20%	30%	40%	50%
Lead Based	**0.47**	**0.55**	0.60	0.70	0.75
MEAD	0.37	0.52	**0.61**	**0.70**	**0.76**
Random	0.25	0.38	0.50	0.58	0.64
Websumm	0.39	0.45	0.53	0.64	0.71

Table 3: English Summaries. Longest Common Subsequence. Average over 10 Clusters. Words and all POS as text representation.

Measure	50W	100W	200W
Cosine (0/1)	0.28	0.28	0.33
Cosine ($tf * idf$)	0.17	0.22	0.43
Word Overlap	0.17	0.17	0.20
Bigram Overlap	0.04	0.04	0.07
Longest Common Subsequence	0.20	0.21	0.23

Table 4: Measures of Similarity between Human Abstract and one Multi-document Summarizer

Measure	50W	100W	200W
Cosine (0/1)	0.33	0.32	0.33
Cosine ($tf * idf$)	0.36	0.44	0.50
Word Overlap	0.20	0.19	0.20
Bigram Overlap	0.06	0.07	0.08
Longest Common Subsequence	0.23	0.25	0.25

Table 5: Measures of Similarity between Human Abstract and one Multi-document Human Extracts

cording to these measures. This experiment shows the use of our framework for comparing human and automatic extracts with human *abstracts*, i.e. coherent, newly written summaries of the documents rather than sentence extracts. To our knowledge, no systematic experiments about agreement on the task of summary writing have been performed before. We believe that our metrics are very valuable, as the highest-quality automatic summaries of the future will probably mirror more and more human summaries, and move away from sentence extracts.

5 Conclusions

In this paper, we have presented a framework for the evaluation of text summarization systems. The contributions of our research are as follows:

First, we provide data and a test-bed for text-summarization evaluation, namely annotation of a pre-existing parallel corpus, human-provided sentence-level utility-judgements which allow us to compile human-generated 'ideal' extracts, and multi-document human-written summaries at different compression rates, following DUC guidelines. These resources are being made available for the community.

Second, we have implemented content-based similarity measures that can be used in both extractive and non-extractive summarization, single and multi-document summarization, and which can be used to compare texts in English and Chinese, and we have shown the advantages of these measures over single co-selection meatures.

Finally, we believe this is the first meta-evaluation directly comparing evaluation measures for text summarization on a large-scale level with unrestricted text.

Acknowledgements

We are grateful to Arda Celebi, John Blitzer, Hong Qi, Danyu Liu, and Elliot Drabek for their work during the workshop. We thank Fred Jelinek, Sanjeev Khudanpur and the staff of the Center for Language and Speech Processing, Johns Hopkins University for their hospitality. We are grateful to Inderjeet Mani. We also thank Chin-Yew Lin, Greg Silber, and Regina Barzilay. The 2001 Summer Workshop at Johns Hopkins University was sponsored by the National Science Foundation via Grant No. IIS-0097467, which included support from the Defense Advanced Research Projects Agency.

A Chinese Extracts

Figures 5, 6, and 7 show Chinese extracts produced by our aligned-based summarization system from the English "human extracts" shown in Section 2.5. Figure 8 shows the summary produced by Mead.

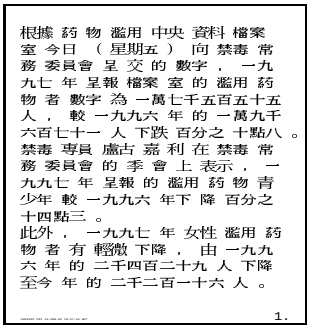

根據 藥 物 濫用 中央 資料 檔案
室 今日 （ 星期五 ） 向 禁毒 常
務 委員會 呈 交 的 數字 ， 一九
九七 年 呈報 檔案 室 的 濫用 藥
物 者 數字 為 一萬七千五百五十五
人 ， 較 一九九六 年 的 一萬九千
六百七十一 人 下跌 百分之 十點八 。
禁毒 專員 盧古 嘉 利 在 禁毒 常
務 委員會 的 季 會 上 表示 ， 一
九九七 年 呈報 的 濫用 藥 物 青
少年 較 一九九六 年下 降 百分之
十四點三 。
此外 ， 一九九七 年 女性 濫用 藥
物 者 有 輕微 下降 ， 由 一九九
六 年 的 二千四百二十九 人 下降
至今 年 的 二千二百一十六 人 。
1.

Figure 5: Aligned-extract by Judge 0

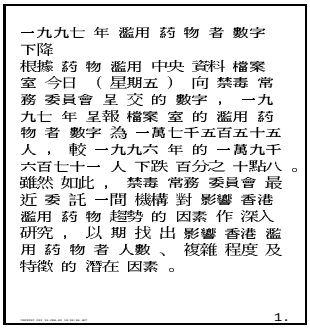
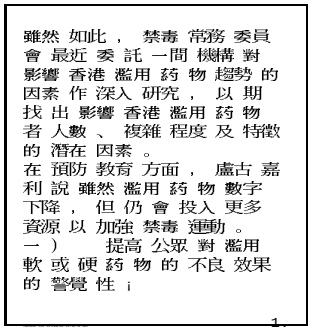

一九九七 年 濫用 藥 物 者 數字
下降
根據 藥 物 濫用 中央 資料 檔案
室 今日 （ 星期五 ） 向 禁毒 常
務 委員會 呈 交 的 數字 ， 一九
九七 年 呈報 檔案 室 的 濫用 藥
物 者 數字 為 一萬七千五百五十五
人 ， 較 一九九六 年 的 一萬九千
六百七十一 人 下跌 百分之 十點八 。
雖然 如此 ， 禁毒 常務 委員會 最
近 委 託 一間 機構 對 影響 香港
濫用 藥 物 趨勢 的 因素 作 深入
研究 ， 以 期 找 出 影響 香港 濫
用 藥 物 者 人數 、 複雜 程度 及
特徵 的 潛在 因素 。
1.

Figure 6: Aligned-extract by Judge 1

References

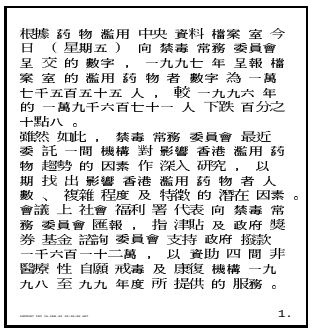
M. Crochemore and W. Rytter. 1994. *Text Algorithms*. Oxford University Press.

R.L. Donaway, K.W. Drummey, and L.A. Mather. 2000. A Comparison of Rankings Produced by


雖然 如此 ， 禁毒 常務 委員 會 最近 委 託 一間 機構 對 影響 香港 濫用 葯 物 趨勢 的 因素 作 深入 研究 ， 以 期 找 出 影響 香港 濫用 葯 物 者 人數 、 複雜 程度 及 特徵 的 潛在 因素 。
在 預防 教育 方面 ， 盧古 嘉 利 說 雖然 濫用 葯 物 數字 下降 ， 但 仍 會 投入 更多 資源 以 加強 禁毒 運動 。
一 ） 提高 公眾 對 濫用 軟 或 硬 葯 物 的 不良 效果 的 警覺 性 ；

1.

Figure 7: Aligned-extract by Judge 2


根據 葯 物 濫用 中央 資料 檔案 室 今 日 （ 星期五 ） 向 禁毒 常務 委員會 呈 交 的 數字 ， 一九九七 年 呈報 檔 案 室 的 濫用 葯 物 者 數字 為 一萬 七千五百五十五 人 ， 較 一九九六 年 的 一萬九千六百七十一 人 下跌 百分之 十點八 。
雖然 如此 ， 禁毒 常務 委員會 最近 委 託 一間 機構 對 影響 香港 濫用 葯 物 趨勢 的 因素 作 深入 研究 ， 以 期 找 出 影響 香港 濫用 葯 物 者 人 數 、 複雜 程度 及 特徵 的 潛在 因素 。
會議 上 社會 福利 署 代表 向 禁毒 常 務 委員會 匯報 ， 指 津貼 及 政府 獎 券 基金 諮詢 委員會 支持 政府 撥款 一千六百一十二萬 ， 以 資助 四 間 非 醫療 性 自願 戒毒 及 康復 機構 一九 九八 至 九九 年度 所 提供 的 服務 。

1.

Figure 8: Chinese Extract by Mead

Summarization Evaluation Measures. In *Proceedings of the Workshop on Automatic Summarization, ANLP-NAACL2000*, pages 69–78. Association for Computational Linguistics, 30 April 2000.

DUC. 2000. *Document Understanding Conference.*

T. Firmin and M.J. Chrzanowski. 1999. An Evaluation of Automatic Text Summarization Systems. In I. Mani and M.T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 325–336. The MIT Press.

W.A. Gale and K.W. Church. 1991. A Program for Aligning Sentences in Bilingual Corpus. In *COLIN 91*, pages 177–184.

C. Grover, C. Matheson, A. Mikheev, and M. Moens. 2000. LT TTT: A Flexible Tokenisation Tool. In *Proceedings of LREC'00.*

E. Hovy and C-Y. Lin. 1999. Automated Text Summarization in SUMMARIST. In I. Mani and M.T. Maybury, editors, *Advances in Automatic Text Summarization*, pages 81–94. The MIT Press.

Hongyan Jing and Kathleen McKeown. 2000. Cut and Paste Based Text Summarization. In *Proceedings of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 178–185, Seattle, Washington, USA, April 29 - May 4, April.

P.A. Jones and C.D. Paice. 1992. A 'select and generate' approach to to automatic abstracting. In A.M. McEnry and C.D. Paice, editors, *Proceedings of the 14th British Computer Society Information Retrieval Colloquium*, pages 151–154. Springer Verlag.

Inderjeet Mani and Eric Bloedorn. 1999. Summarizing similarities and differences among related documents. *Information Retrieval*, 1(1):35–67.

I. Mani, D. House, G. Klein, L. Hirshman, L. Obrst, T. Firmin, M. Chrzanowski, and B. Sundheim. 1998. The TIPSTER SUMMAC Text Summarization Evaluation. Technical Report Technical Report MTR 98W0000138, The Mitre Corporation, McLean, Virginia.

I. Mani, T. Firmin, and B. Sundheim. 2001. Summac: A text summarization evaluation. *Natural Language Engineering.*

A. Mikheev. 2000. Tagging Sentence Boundaries. In *Proceedings of the NAACL*, Seattle, USA. ACL.

C.D. Paice and M.P. Oakes. 1999. A Concept-Based Method for Automatic Abstracting. Technical Report Research Report 27, Library and Information Commission.

Dragomir R. Radev, Hongyan Jing, and Malgorzata Budzikowska. 2000. Centroid-based summarization of multiple documents: sentence extraction, utility-based evaluation, and user studies. In *ANLP/NAACL Workshop on Summarization*, Seattle, WA, April.

Horacio Saggion and G. Lapalme. 2000. Concept Identification and Presentation in the Context of Technical Text Summarization. In *Proceedings of the Workshop on Automatic Summarization. ANLP-NAACL2000*, Seattle, WA, USA, 30 April. Association for Computational Linguistics.

Horacio Saggion, Dragomir Radev, Simone Teufel, Lam Wai, and Stephanie Strassel. 2002. Developing Infrastructure for the Evaluation of Single and Multi-document Summarization Systems in a Cross-lingual Environment. In *Proceedings of LREC 2002: Language Resources and Evaluation Conference*, volume II, pages 747–754, Las Palmas, Spain, 29-31 May. ELRA.

Gerard Salton, James Allan, Chris Buckley, and Amit Singhal. 1994. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, 264:1421–1426.

G. Salton. 1988. *Automatic Text Processing.* Addison-Wesley Publishing Company.

K. Sparck Jones and J.R. Galliers. 1995. *Evaluating Natural Language Processing Systems: An Analysis and Review.* Number 1083 in Lecture Notes in Artificial Intelligence. Springer.