# News Abridgement Algorithm Based on Word Alignment and Syntactic Parsing[1]

Min Yu[1], Huaping Zhang[1], Yu Zhang[1], Yang Qiao[1],
Zhonghua Zhao[2], Yueying He[2]
[1] Beijing Institute of Technology School of Computer Science, Beijing, 100081, China
[2] National Computer Network and Information Security Management Center, Beijing, 100031, China
yumin2014@nlpir.org, kevinzhang@bit.edu.cn, zhangyu2014@nlpir.org,
qiaoyang2014@nlpir.org, zhaozh@cert.org.cn, hyy@cert.org.cn

**Abstract:** The rapid development of new media results in a lot of redundant information, increasing the difficulty of quickly obtaining useful information and browsing simplified messages on portable devices. Thus emerges the automatic news abridgement technology. We propose a novel method of word alignment, aiming at news headlines, applying the combination method of statistics and rules to intelligent abridgement. And a new framework based on the combination of sentence abridgement and sentence selection to generate the abridgement result of news contents, abridging the original text to the word limit, in order to achieve the uttermost conservation of the original meaning. Meanwhile, for a fair and intelligent evaluation, this paper presents an evaluation method of automatic summarization specific to sentence abridgement techniques. Experimental results show that the proposed methods are feasible, and able to automatically generate coherent and representative summaries of given news with high density.

**Key Words:** News Abridgement； Keyword Features；Sentence Abridgement；Heuristic Rules；Sentence Selection

## 1 Introduction

With the rapid development of Internet and digital technology, a variety of new media which supports users in interaction has sprung up. But media contents are fragmented and irregular, making it difficult to select information [1]. To help readers find high-quality content from a large quantity of news, as well as to display the news title lists on small handheld screens, we need to abridge and filter the text contents of Internet information for better representation. In this context, some tools are urgently needed to find desired information automatically, quickly and intelligently from massive amounts of news.

Most of current researches extract main contents of news as the summaries of the original news, with no regard for the abridgement of news titles and contents, making the word limit problem especially prominent. The traditional manually method of news abridgement is time-consuming when faced with massive media information data, thus unable to meet the demand for real-time processing of the various requests.

Based on these problems and challenges, this paper presents a novel algorithm based on word alignment technology and syntactic analysis to abridge news. Experimental results show that the

[1] Min Yu is the corresponding author. (E-mail: yumin2014@nlpir.org)

algorithm is feasible, and able to automatically generate a coherent and representative summary of the given news with high density.

The contributions of our paper are as follows:

1) A novel word alignment method is applied into news title abridgement, which integrates application rules and statistical methods to achieve intelligent abridgement.

2) A framework that combines sentence abridgement and sentence selection is proposed to generate abridgement results of news content.

3) An automatic summarization evaluation method specific to sentence abridgement is proposed.

## 2 Related Work

Most current methods abridge sentences by removing unimportant words or phrases with supervised and unsupervised learning algorithms [2]. Knight and Marcu [3] model the sentence abridgement process with noise channel model, and propose an abridgement algorithm based on Decision Tree Learning. Some studies use supervised learning algorithms such as CRFs [4], support vector regression [5] and so on. Supervised methods can easily integrate various features, but fail to deal with global dependencies due to computational efficiency and other reasons. Unsupervised methods are mainly based on integer linear programming [6] [7], which struggle with integrating various features, while handle global dependencies well.

Abstract generation mainly contains two methods: extraction and summarization [8]. Extraction approach focuses on selecting important sentences from the document and coupling them into a new summary, or summarizing with the formulated optimization framework, including integer linear programming (ILP) and sub-module functions. Gillick et. al. [9] put forward an ILP method based on concepts. Li et. al. [10] assess concepts weights of ILP structure by supervision strategies. Summarization methods are considered more difficult, which involve in-depth content like sematic representation, content arrangement and surface representation, and require complex techniques like natural language processing.

## 3 News Abridgement Algorithm Based On Word Alignment and Syntax Parsing

This section describes the interrelated algorithms of news abridgement, including sentence abridgement algorithm based on word alignment, word weighting based on features combination, sentence weighting based on features combination, heuristic sentence abridgement algorithm based on keywords. These algorithms build up a news abridgement system. Each algorithm will be described in detail below.

### 3.1 Sentence Abridgement Algorithm Based on Word Alignment

To shorten news titles, we introduce the "Tongyici Cilin" Extensions [11] as semantic dictionary to replace the original words with shorter synonyms which are called alignment words in this paper. With the aim of identifying the short words and new words which are appropriated to current network vocabulary from many synonyms, we use the edit distance algorithm to align the words with different

meanings and obtain the corresponding transition probability, based on which the sentence patterns and word usage when news titles are written can be acquired using maximum probability principle. Ultimately the system generates novel title abridgement under the word limit, maintaining the original meaning.

In this paper, transition probability matrix is obtained by monolingual word alignment [12] [13] and showed in table 1.

**Table 1.** Examples of the transition probability of the original word to the target word

| |
|---|
| 录用（NULL，0.33；录用，0.67） |
| 【employ (NULL, 0.33; employ, 0.67)】 |
| 中央巡视组（中央巡视组，0.25；NULL，0.63；整改，0.13） |
| 【Central inspection team (Central inspection team, 0.25; NULL, 0.63; rectify, 0.13)】 |
| 什么样（NULL，0.25；什么样，0.75） |
| 【what kind (NULL, 0.25; what kind, 0.75)】 |
| 干部（NULL，0.16；领导，0.01；干部，0.82） |
| 【cadre (NULL, 0.16; leader, 0.01; cadre, 0.82)】 |
| 山体（NULL，0.07；山体，0.93） |
| 【mountain (NULL, 0.07; mountain, 0.93)】 |
| 小学（NULL，0.09；小学，0.91） |
| 【primary school (NULL, 0.09; primary school, 0.91)】 |

In the table above, the alignment words which can replace the original words and their corresponding probability are bracketed. For example, 录用（NULL，0.33；录用，0.67），【employ (NULL, 0.33; employ, 0.67)】, means in conventional manual processing, the original word "录用（employ）" is replaced by NULL with probability 0.33, is replaced by "录用（employ）" with probability 0.67. We take the alignment word "录用（employ）" with the maximum transition probability as the ultimate alignment word. Thus, compared to the method using only the "Tongyici Cilin" for word alignment, the method we propose is more statistically adaptive to corpus and is able to handle large-scale corpus of news.

### 3.2 Word Weighting Based on Features Combination

We put forward an algorithm based on multi-feature of sentence and prolixity processing combined with the heuristic sentence abridgement algorithm based on keywords to abridge news text.

Many factors are to be considered in news text abridgement, including word frequency, part of speech, word length and position of word. We integrate all factors and put forward a method based on features combination to calculate the weight of words in news text.

**1) Position of word**

Words appearing in the title or some other positions are considered important. The first sentence of news usually introduces its five elements: when, where, who, why, and what is going on.

We identify the five elements of news from the first sentence by named entity recognition [14], and give weights by the following principles.

$$Loc(w_i)=\begin{cases}1, & \text{word } w_i \text{ is one of the word in news title or the five elements;}\\ \beta, & \text{word } w_i \text{ appears in the second sentence or the end of paragraph } (0<\beta<1);\\ 0, & \text{other position}\end{cases} \quad (1)$$

**2) Keyword**

We extract the key features from news content of data sets with bi-directional matching algorithm [15] and obtain 2105 keywords, which are given higher weights.

$$Key(w_i)=\begin{cases}1, & word\ w_i\ is\ in\ the\ key\ word\ list;\\ 0, & other\ case\end{cases} \quad (2)$$

The keyword list is obtained through machine learning and adding new words. Some results are shown in table 2.

**Table 2.** Examples of keywords extraction results

| keywords | keywords | keywords |
|---|---|---|
| 高新技术<br>（high technology）<br>三中全会<br>（Third Plenary Session）<br>无期徒刑<br>（life imprisonment）<br>易燃易爆<br>（inflammable and explosive）<br>公款吃喝<br>（public money eating and drinking）<br>范剑平<br>（Fan Jianping）<br>常住人口<br>（permanent population）<br>羊城晚报<br>（Yangcheng Evening News）<br>一二线城市<br>（a second tier cities）<br>新闻发布会<br>（press conference） | 率先垂范<br>（the first example）<br>觅仙泉<br>（Mi Xianquan）<br>严雪花<br>（Yan Xuehua）<br>达赖喇嘛<br>（Darai Lama）<br>女德班<br>（women in Durban）<br>阳宝华<br>（Yang Baohua）<br>强盗逻辑<br>（gangster logic）<br>终期考核<br>（final examination） | 杨六斤<br>（Yang Liujin）<br>警钟长鸣<br>（keep ringing the alarm bell）<br>黄某峰<br>（Huang Moufeng）<br>单霁翔<br>（Shan jixiang）<br>总揽全局<br>（overall authority）<br>纪检监察部门<br>（discipline inspection and supervision departments）<br>革委会副主任<br>（Deputy director of the Committee） |

**3) Part of speech**

Nouns and compound nouns play important roles in expressing the meaning of articles.

$$POS(w_i)=\begin{cases}1, & word\ w_i\ is\ compound\ noun;\\ \gamma, & word\ w_i\ is\ the\ name\ of\ people\ or\ place\ or\ organization\ (0<\gamma<1);\\ 0, & other\ part\ of\ speech\end{cases} \quad (3)$$

Combining these three feature formulas, we design a method based on feature combination to calculate the weight of words.

$$Score(w_i) = n_i + \lambda * (Loc(w_i) + Key(w_i) + POS(w_i)) \quad (4)$$

Where $n_i$ is the number of times that word $w_i$ has appeared, and the relative optimal values of the parameters $\beta$、$\gamma$ and $\lambda$ are determined according to the repeated adjustment of the experiment. Meanwhile, the value of $\gamma$ is related to the length of article, and the empirical value is 15.

The weight of word $w_i$ calculated with formula (4) is as follows,

$$W(w_i) = Score(w_i) / \max_{w_j \in d}\{Score(w_j)\} \quad (5)$$

Where d is document, x is the other words in the document.

### 3.3 Sentence Weighting Based on Features Combination

In order to give higher weights to the sentences that express the topic of document, this paper defines a series of features to weigh the importance of each sentence.

**1) Content of sentence**

The more words and phrases with high weight the sentence contains, the greater the amount of information is, and the more important the sentence is.

$$Cont(s_i) = \sum_{j=1}^{N} \sqrt{W(w_j)} / N \quad (6)$$

Where N is the number of words in sentence $s_i$, word $w_j \in s_i$, $0 < Cont(s_i) \leq 1$.

**2) Position of sentence**

We give the corresponding weight to a sentence according to where it appears, and get the formula (7).

$$Loc(s_i) = \begin{cases} 1, & Sentence\ s_i\ is\ the\ first\ sentence\ of\ the\ paragraph; \\ 0.5, & Sentence\ s_i\ is\ the\ last\ sentence\ or\ the\ second\ sentence\ of\ the\ paragraph; \\ 0, & other\ position \end{cases} \quad (7)$$

Considering the content and position of the sentence, a weighted linear combination of multiple features is the final weight of the sentence. The formula for calculating the weight of a sentence is as follows:

$$W(s_i) = \varphi * Cont(s_i) + \eta * Loc(s_i) \quad (8)$$

Where $\varphi$ and $\eta$ are adjustable parameters, $\varphi + \eta = 1$.

### 3.4 Heuristic Sentence Abridgement Algorithm Based on Keywords

This algorithm applies heuristic abridgement rule to the results of the syntactic parsing of each sentence [16], and determines whether the sentence is to be removed or retained by the composition of each node in the syntax tree and the keyword feature. In this approach, we combine multiple constraints to improve the linguistic quality of abridged sentences [17].

We obtain a set of rules and the weights of rules through machine learning. For example, "remove the contents in front of the first noun phrase", "remove the adverbs or adjective phrase in a sentence", "remove the preposition phrases as attributive or adverbial in a sentence " and so on.

It is also important to determine the weights in abridgement rules. Through machine learning, it is found that keywords being given higher weights in the abridgement rules ensures the soundness of the abridgement process. Specifically, the weights of keywords in news titles and news texts are assigned to 3 and 2, and the weights of other words in the sentence are assigned to 0. For word that meets the abridgement rules, its weight will be reduced by 1 in each loop. If the grammar requirement of the abridged sentence is high and the length of the abridged sentence is not required, the loop can be ended. And an abridged sentence is formed with all the words with non-negative weights. If there is a requirement for the length of abridged sentence, constraints need to be added to the sentence selection method based on integer linear programming.

To restore the grammar and semantic of abridged sentence, we use the vector space model to calculate the sentence similarity of the pre-extracted abstract to remove the redundancy in the abstract [18], and finally use sentence selection based on ILP to generate the abridgement of news.

## 4 Experiment and Evaluation

This section is divided into two parts, the experiment of news title abridgement and the news text abridgement.

### 4.1 News Title Abridgement Experiment

The corpus of news title abridgement algorithm contains 5658 pairs of news titles extracted from people.cn Web (original titles) and Client (target titles). This algorithm is based on the transition probability mentioned in section 3.1. In the experiment, 5558 title pairs are chosen as training set, which are manually labeled, with the left 100 pairs as test set. In order to make this model have wide range of adaptability, the selection of sentences is assessed against the topics, lengths and syntactic structure components.

There are three criteria in traditional sentence abridgement evaluation: the importance of words, grammar normalization, and compression ratio. By analyzing the criteria above, a comprehensive evaluation of sentence abridgement function is given in equation (9):

$$\text{Score} = \text{Gram} + \text{Impo} + 5\times(1 - \text{CompRate}) \quad (9)$$

Where Gram represents normative grammar, Impo indicates the importance of words, and CompRate means the compression ratio, so that Score denotes sentence abridgement overall score. Through evaluation, the compression ratio is automatically determined by the system, semantics and grammar normalization are evaluated manually. Evaluation results are shown in table 4.

Several typical examples of abridgement results are shown in table 3, and are elaborated as follows.

**Table 3.** Examples of artificial abridgement and system abridgement for news titles

| | | |
|---|---|---|
| Example 1 | original title | 35 城市一卡通实现互联互通异地刷卡<br>（35 cities achieve off-site and interconnected city card） |
| | artificial abridgement | 35 个城市一卡通实现异地刷卡<br>（35 cities achieve off-site city card） |
| | system abridgement | 35 城市一卡通实现异地刷卡<br>（35 cities achieve off-site city card） |
| Example 2 | original title | 韩疗养医院火灾系八旬患者放火引起<br>(The fire happening in a Korean nursing home was set by a eighty-year-old patient) |
| | artificial abridgement | 韩疗养院火灾系八旬患者放火<br>(The fire happening in a Korean nursing home was set by a eighty-year-old patient) |
| | system abridgement | 韩医院火灾系八旬患者放火<br>(The fire happening in a Korean nursing home was set by a eighty-year-old patient) |

As table 3 shows, the manual sentence compression ratio of example 1 is 82.35%, whereas the automatic compression ratio is 76.47%. Comparing with manual result, automatic result not only in the grammatical and semantic conforms to the standard, but also uses fewer words to express the core meaning of original sentence. As for example 2, the system abridged sentences are not precise to express the original meaning, by matching synonym "韩疗养院（Korean nursing home ）" as "韩医院（Korean hospital）", however the main meaning remains the same, also the result is within the acceptable range.

**Table 4.** News title abridgement system evaluation results

| | grammar | semantics | compression ratio |
|---|---|---|---|
| system abridgement | 68.8% | 69.6% | 75.6% |
| artificial abridgement | 88.8% | 77.6% | 72.8% |
| system / artificial approximation | 77.48% | 89.69% | 103.85% |

By observing the results presented in table 4, the overall system output results are inferior to manual abridgement results, but all of the three aspects of system abridgement results are above 60%. Moreover, in terms of grammar and semantics, system output basically conforms to the specification of Chinese. This also illustrates that the system has advantages in maintaining the syntactic structure and semantic content of the abridged sentences. In addition, the compression ratio of the system is slightly lower, and is close to manual result, therefore the system preliminarily achieves the goal of abridging titles.

### 4.2 Experiment of News Text Abridgement

100 news articles from the people.cn Web are selected as the experimental data to do abstract extraction and analysis. Similarly, they are assessed against topics, lengths, and syntactic structures, which guarantees the generalization of the model, so as to better evaluate its performance.

**1) Keywords extraction experiment results**

Due to space limitations, part of results is given in Table 2.

**2) News content abstract pre-extraction**

According to the results of repeated testing procedures, the parameters $\beta$、$\gamma$ and $\lambda$ in the formulas (1), (3) and (4) are valued 0.5, 0.5 and 15 respectively, and the parameters $\varphi$ and $\eta$ in the formula (8) are both assigned as 0.5. The weights of sentences are calculated according to their multiple features , and then take the top 15 sentences as the result of the pre-extraction of the news abstract. Next, these sentences are sorted based on their relative locations in the original document.

**3) Pre-extraction Abstract sentence abridgement**

Pre-extracted sentences abridgement and restoration are based on heuristic linguistic rules. The results of evaluation are given in Table 5.

**Table 5.** Pre-extraction abstract sentence abridgement result evaluation

| grammar | semantics | compression ratio |
|---|---|---|
| 77.93% | 80.89% | 73.15% |

As it shows, the semantic result is higher than the result of grammar, for that the abridgement rules might cause incomplete sentences, and so reduce the grammatical characteristics of the sentence. However, since the abridgement algorithm is based on keywords, it still retains the important information of the sentence, so that obtain a relatively higher semantic score.

**4) News abridgement system evaluation**

Commonly used evaluation method is assuming there are x sentences in the standard summary, y sentences in the generated summary, and K sentences that appear in both the standard summary and the generated summary. Therefore, the precision and recall are $\mathrm{P}=\frac{\mathrm{k}}{\mathrm{x}}$ and $\mathrm{R}=\frac{\mathrm{k}}{\mathrm{y}}$. $\mathrm{F}_1-\mathrm{score}$ is a compromised evaluation index considering precision and recall, $\mathrm{F}_1-\mathrm{score}=\frac{2\times\mathrm{P}\times\mathrm{R}}{\mathrm{P}+\mathrm{R}}$. Since the extraction task of this paper involves sentence abridgement, which may cause k equals to 0, we propose a novel definition for precision and recall.

$$P=\frac{k}{x}\times((\sum_{i=1}^{k}Score)/k) \quad (10)$$

$$R=\frac{k}{y}\times((\sum_{i=1}^{k}Score)/k) \quad (11)$$

Where k denotes the number of similar sentences. And the results of precision, recall and $\mathrm{F}_1-\mathrm{score}$ are shown in table 6.

**Table 6:** News abridgement system evaluation

| precision | recall | $F_1 - score$ |
|---|---|---|
| 79.26% | 76.32% | 77.76% |

The experimental results suggest that the news abridgement algorithm is feasible and the extraction of abstract utilizes more fine-grained approach and are not just to sentence level extraction; for given news articles, it could automatically generate coherent and representative news summaries with high density.

An news abridgement example by using the proposed algorithm is given below (http://www.chinanews.com/gj/2014/05-27/6214377.shtml). The original news has in total 1303 words, and abridged news contains only 190 words, and it not only satisfies the grammar and semantic requirements but also significantly reduces the time of consideration in manual abridgement.

Abridged news title：

莫迪就任印度第 15 任总理(Modi became the fifteenth Prime Minister of India)

Abridged news text：

外媒 27 日报道，印度人民党党首莫迪就任成为印度共和国第 15 任总理，承诺建立一个“强大和具包容性”的印度，且组建了一个人数少很多的精简内阁。4000 名宾客见证了莫迪的就职，包括印度的巴基斯坦总理谢里夫。“我们创造辉煌的未来……让我们梦想，建立一个强大、发达和具包容性的印度，印度合作，推动世界和平与发展。”印度人民党主席拉杰纳特·辛赫任内政部长。莫迪被认为具有“亲市场”倾向，印度经济界有期待。(As foreign media reported on 27th, the leader of the Bharatiya Janata Party (BJP) Modi has become the 15th Prime Minister of the Republic of India, committed to establishing a "strong and inclusive" India and setting up a compact minister council. 4000 guests witnessed the inauguration of Modi, including India's Pakistani Prime Minister Nawaz Sharif. "We will create a brilliant future..... Let us dream, build a strong, developed and inclusive India, India will cooperate on promoting peace and development of the world." Rajnath, chairman of India people's party, he served as Minister of the interior. Modi is considered to have pro-market tendency, and India economic community have expectations.)

The experimental results show that the automatic news text abridgement system proposed in this paper achieves good results.

## Conclusion

As the experiment results suggest, the proposed news abridgement algorithm is feasible. The algorithm could automatically generate coherent and representative summaries of the given news with high density, which could significantly reduce the workload of news editing and abridgement.

## References

1. Zhang Jing. Research on the characteristics and problems of network entertainment news headlines [J]. Journalism Knowledge, 2011，11: 110-111.

2. Jing H. Sentence reduction for automatic text summarization[C]. Proceedings of the sixth conference on Applied natural language processing. Association for Computational Linguistics, 2000：310-315.
3. Knight K, Marcu D. Summarization beyond sentence extraction: A probabilistic approach to sentence compression[J]. Artificial Intelligence, 2002, 139(1) : 91-107.
4. Nomoto T. A comparison of model free versus model intensive approaches to sentence compression[C]. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing：Volume 1-Volume 1. Association for Computational Linguistics, 2009：391-399.
5. Galanis D, Androutsopoulos I. An extractive supervised two-stage method for sentence compression[C]. Human Language Technologies：The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2010：885-893.
6. Filippova K, Strube M. Dependency tree based sentence compression[C]. Proceedings of the Fifth International Natural Language Generation Conference. Association for Computational Linguistics, 2008：25-32.
7. Clarke J, Lapata M. Global inference for sentence compression：An integer linear programming approach[J]. Journal of Artificial Intelligence Research, 2008：399-429.
8. Li C, Liu F, Weng F, et al. Document Summarization via Guided Sentence Compression [C]. EMNLP. 2013:490-500.
9. Dan Gillick, Benoit Favre, Dilek Hakkani-Tur, et al.The icsi/utd summarization system at tac 2009[C]. In Proceedings of TAC, 2009.
10. Li C, Qian X, Liu Y. Using Supervised Bigram-based ILP for Extractive Summarization[C]. ACL (1). 2013: 1004-1013.
11. "Tongyici Cilin" Extensions. http：//www.ir-lab.org/
12. Och F J, Ney H. A comparison of alignment models for statistical machine translation[C]. Proceedings of the 18th conference on Computational linguistics-Volume 2. Association for Computational Linguistics, 2000: 1086-1090.
13. Yarowsky D, Wicentowski R. Minimally supervised morphological analysis by multimodal alignment[C]. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2000: 207-216.
14. Nadeau D, Sekine S. A survey of named entity recognition and classification[J]. Lingvisticae Investigationes, 2007, 30(1): 3-26.
15. Zhang Ruiqi. Research on Top N Hot Topics Detection Method Based on Key Features Clustering[D]. Beijing：Beijing Institute of Technology，2015.
16. http：//nlp.stanford.edu/software/lex-parser.shtml.
17. Zajic D, Dorr B J, Lin J, et al. Multi-candidate reduction: Sentence compression as a tool for document summarization tasks[J]. Information Processing & Management, 2007, 43(6): 1549-1570.
18. Zhou, Guangyou, et al. Towards faster and better retrieval models for question search. Proceedings of the 22nd ACM international conference on Conference on information & knowledge management. ACM, 2013.