VSoLSCSum: Building a Vietnamese Sentence-Comment Dataset for Social Context Summarization

Minh-Tien Nguyen[1,2], **Viet Dac Lai**[1], **Phong-Khac Do**[1], **Duc-Vu Tran**[1], **and Minh-Le Nguyen**[1]
[1] School of Information Science,
Japan Advanced Institute of Science and Technology (JAIST), Japan.
[2] Hung Yen University of Technology and Education, Vietnam.
{tiennm, vietld, phongdk, vu.tran, nguyenml}@jaist.ac.jp

Abstract

This paper presents *VSoLSCSum*, a Vietnamese linked sentence-comment dataset, which was manually created to treat the lack of standard corpora for social context summarization in Vietnamese. The dataset was collected through the keywords of 141 Web documents in 12 special events, which were mentioned on Vietnamese Web pages. Social users were asked to involve in creating standard references and the label of each sentence or comment. The inter-agreement calculated by Cohen's Kappa among raters after validating is 0.685. To illustrate the potential use of our dataset, a learning to rank method was trained by using a set of local and cross features. Experimental results indicate that the summary model trained on our dataset outperforms state-of-the-art baselines in both ROUGE-1 and ROUGE-2 in social context summarization.

1 Introduction

In the context of social media, users can freely discuss the content of an event mentioned in a Web document in the form of comments. For example, after reading an event, e.g. CASA rescue airplane explosion from Dan Tri[1], readers can write their comments on the interface of Dan Tri. These comments, one form of social information (Amitay and Paris, 2000; Delort et al., 2003; Sun et al., 2005; Hu et al., 2008; Lu et al., 2009; Yang et al., 2011; Wei and Gao, 2014; Nguyen and Nguyen, 2016), have two critical characteristics: (i) reflecting the content and sharing the topic of a Web document, and (ii) revealing the opinions of readers with respect to that event. This observation inspires a novel summarization task, which utilizes the social information of a Web document to support sentences for generating summaries.

Automatic summarization was first studied by (Luhn, 1958; Edmundson, 1969). Until now, extractive summarization methods usually focus on plain-text documents and select salient sentences by using statistical or linguistic information in the form of binary classification (Kupiec et al., 1995; Conroy and O'Leary, 2001; Osborne, 2002; Yeh et al., 2005; Shen et al., 2007; Yang et al., 2011; Cao et al., 2015b). These methods, however, only consider internal information of a Web document, e.g. sentences while ignoring its social information.

Social context summarization is a task which selects both important sentences and representative comments from readers of a Web document. It has been studied by using different kind of social information such as hyperlinks (Amitay and Paris, 2000; Delort et al., 2003), click-through data (Sun et al., 2005), comments (Delort, 2006; Hu et al., 2007; Hu et al., 2008; Lu et al., 2009), opinionated text (Kim and Zhai, 2009; Ganesan et al., 2010; Paul et al., 2010), or tweets (Yang et al., 2011; Gao et al., 2012; Wei and Gao, 2014; Wei and Gao, 2015; Nguyen and Nguyen, 2016). (Yang et al., 2011) proposed a dual wing factor graph model for incorporating tweets into the summarization and used Support Vector Machines (SVM) (Cortes and Vapnik, 1995) and Conditional Random Fields (CRF) (Lafferty et al., 2001) as preliminary steps in calculating the weight of edges for building the graph. (Wei and Gao, 2014) used a learning to rank (L2R) approach with 35 features trained by RankBoost for news highlight extraction. (Nguyen et al., 2016c) extended the work of (Wei and Gao, 2014) by proposing entailment and

[1] http://dantri.com.vn

Proceedings of the 12th Workshop on Asian Language Resources,
pages 38–48, Osaka, Japan, December 12 2016.

semantic features for summarizing Web documents and their comments. In contrast, (Gao et al., 2012) proposed a cross-collection topic-aspect modeling (cc-TAM) as a preliminary step to generate a bipartite graph, which was used by a co-ranking method to select sentences and tweets for multi-document summarization. (Wei and Gao, 2015) proposed a variation of LexRank, which used auxiliary tweets for building a heterogeneous graph random walk (HGRW) to summarize single documents. (Nguyen and Nguyen, 2016) proposed SoRTESum, a ranking method using a set of recognizing textual entailment features (Nguyen et al., 2015) for single-document summarization. However, these methods were applied for English. To the best our knowledge, no existing method studies social context summarization for Vietnamese due to the lack of a standard corpora.

The objective of this study is to create a standard corpus for social context summarization in Vietnamese. This paper makes the following contributions:

- We create and release a Vietnamese dataset[2] which can be used to evaluate summary methods in social context and traditional summarization. The dataset includes 141 Web documents with their comments in 12 special events. The gold-standard references are selected by social users.
- We investigate social context summarization by state-of-the-art summary approaches. This investigation helps to point out the best summarization method in this task. Our demo system can be also accessed[3].

In the following sections, we first introduce the creation of our dataset with detail observation. Next, we show the formulation of summarization in the form of a learning to rank task. After training a summary model, we compare our results with various summary methods, along with discussion and analysis. We finish by drawing important conclusions.

2 VSoLSCSum for Summarization

This section shows the creation of our dataset in three steps: annotation framework introduction, data collection and data annotation with deep observation, and summarization.

2.1 Annotation Framework

The dataset was created by using a framework shown in Figure 1. The framework contains two main modules: data collection and data annotation. The data collection receives a keyword corresponding to an event, then collects a bunch of Web documents related to this event. Afterward, the pre-processing step eliminates unnecessary information, e.g. HTML, and tokenizes sentences. In the annotation module, raw texts were shown on an annotation website where social users were asked to annotate each document and its comments based on an instruction.

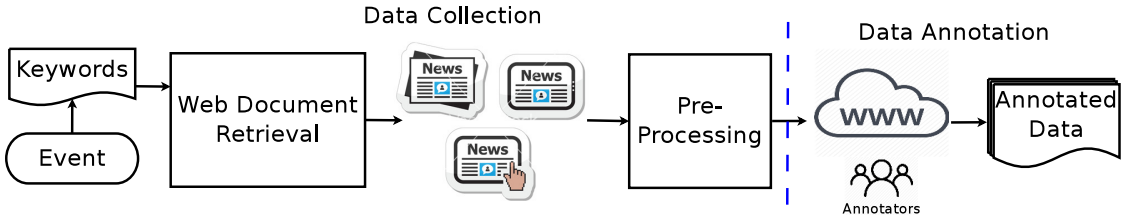


Figure 1: The overview of annotation framework

2.2 Data Collection

To create the dataset, 12 special events appearing on Vietnamese Web pages in September 2016 were first identified. Each event was empirically assigned by a noun phrase keyword which reflects the major object of the event. The noun phrase is a major entity which appears in an event. For example, in CASA rescue airplane explosion, the keyword is *"casa"*. It is possible to define a list of keywords for each

[2]Download at: https://github.com/nguyenlab/VSoLSCSum-Dataset
[3]http://150.65.242.101:9293/?paper=coling_alr

event. However, we collect the Web documents of an event from several news providers (non-duplicate); therefore, creating a set of keywords is unnecessary. All keywords are shown in Table 1.

Table 1: The events and corresponding keywords

Event	Keyword	Event	Keyword
CASA rescue airplane explosion	"*casa*"	Michael Phelps golden medal	"*michael phelps*"
American president election	"*donald trump*"	Pokemon Go game	"*pokemon go*"
Formosa pollution	"*formosa*"	Tan Son Nhat airport	"*tan son nhat*"
Vietnamese Olympic godel medal	"*hoang xuan vinh*"	Trinh Xuan Thanh	"*trinh xuan thanh*"
IS Islamic State	"*is*"	Vu Quang Hai	"*vu quang hai*"
Murder in Yen Bai	"*yen bai*"	Huynh Van Nen	"*huynh van nen*"

After defining keywords, a set of relevant Web documents was retrieved by using HTMLUnit library[4]. Subsequently, raw data was collected by parsing the Web documents using JSOUP parser[5]. The information of each document contains six elements shown in Table 2. In parsing, sentences and comments were also tokenized[6].

Table 2: The elements of a Web document

Element	Description
Title	The title of a Web document
Abstract	A short summary of a Web document written by writer
Content	The content of a Web document
Writer	The writer of a Web document
Comment	A set of comments showing the opinions from readers
Tag	A set of keywords which indicate the topic of a Wed document

The dataset consists of 141 open-domain articles along with 3,760 sentences, 2,448 gold-standard references, and 6,926 comments in 12 events. Note that the gold-standard references also include comments. Table 3 shows the statistics of the dataset.

Table 3: Statistical observation; *s*: sentences, *c*: comments.

Documents	Sentences	Summaries	Comments	Observation	Sentences	Comments
141	3,760	2,448	6,926	# positive examples	1,343	964
# Tokens	83,010	60,953	93,733	# negative examples	2,417	5,962
# Avg-sentences/article	26.666	17.361	49.120	% positive examples	35.718	13.918
# Avg-tokens/article	588.723	432.290	664.773		—	—
# Avg-tokens/sentence	22.077	24.899	13.533	% Token overlapping	s/c: 37.712	c/s: 44.820

2.3 Data Annotation

Data creation was conducted in two steps: annotation and validation. In the annotation, to ask social users, an annotation website was created for annotating this data[7]. Five native Vietnamese speakers involved to annotate the dataset. Each annotator read a complete document and its comments to select summary sentences and comments (called instances) which reflect the content of each document. Each sentence or comment was assigned a Cosine score calculated by bag-of-words model, which measures the similarity of the abstract and the current sentence or comment. The Cosine score indicates that a summary sentence or comment should include salient information of a Web document mentioned in the abstract. Note that the score is only used to calculate the similarity between sentences with the abstract (or comments with the abstract). In selection stage, annotators have to consider the following constraints to select a summary sentence or comment:

[4]http://htmlunit.sourceforge.net/
[5]https://jsoup.org
[6]http://mim.hus.vnu.edu.vn/phuonglh/softwares/vnTokenizer
[7]http://150.65.242.91:9080/vn-news-sum-annotator/annotate

- Each chosen sentence or comment has to reflect the content of a document.
- The Cosine score of each sentence or comment affects the selection. The higher Cosine score of a sentence or comment is, the higher probability of this sentence or comment should be selected.
- The selected instances are no less than four sentences and six comments (less than 30% of average sentences per document, see Table 3). The total selected instances are no more than 30, including both sentences and comments.

The label of a sentence or comment was generated based on majority voting among social annotators. For example, given a sentence, each annotator makes a binary decision in order to indicate whether this sentence is a summary candidate (YES) or not (NO). If three annotators agree yes, this sentence is labeled by 3. Therefore, the label of each sentence or comment ranges from 1 to 5 (1: very poor, 2: poor, 3: fair, 4: good; 5: perfect). The gold-standard references are those which receive at least three agreements from annotators (3/5).

Table 4: A translated example of label from five annotators, taken from Pokemon Go event.

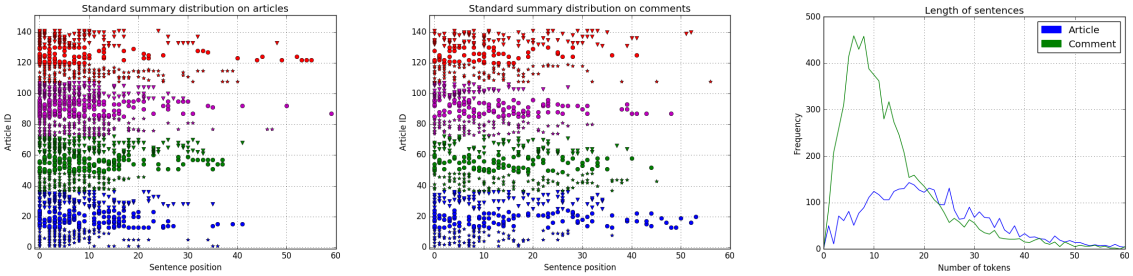
Sentence	Label
This game requires the move of users to search the virtual pet, collect balls and eggs (S)	1:0:1:1:1
The idea of Pokemon Go is a significant improvement (C)	0:1:0:0:0

Table 4 shows a translated example from Vietnamese texts of Pokemon Go event. The sentence (S) receives four agreements over five annotators, so its final label is 4 and it becomes a standard reference. The label of comment (C) is 1 due to only one agreement, then it is non-standard reference.

In the validation, to ensure the quality of the dataset, two other native Vietnamese raters were asked to vote each sentence or comment, which were already labeled. The inter-agreement was calculated based on the voting of the two users. The agreement was computed by Cohen's Kappa[8] between the two annotators is 0.685 with 95% confidence interval. The strength of agreement is considered to be good.

2.4 Data Observation

Table 3 (right table) illustrates two primary points: (i) there exists common words or phrases between sentences and comments (the last right row) and (ii) readers tend to use words or phrases appearing in sentences to create their comments (44.820% of word overlapping of comments on sentences).



(a) Standard sentence references (b) Standard comment references (c) Sentence length

Figure 2: The position of standard summaries over 12 events and sentence length distribution.

The position of gold-standard references and sentence length over the corpus were also observed, in which color points in Figures 2a and 2b represent gold-standard sentences and comments. Figures 2a and 2b show that: (i) gold-standard references locate within first 10 sentences and top 20 comments, and (ii) standard-comment references tend to appear in a wider range compared to sentences. Figure 2c indicates that the length distribution of almost sentences ranges from five to 40 and of almost comments are from three to 40. The average sentence length and comment length is 22.077 and 13.533 respectively (see Table 3).

[8]https://graphpad.com/quickcalcs/kappa1/

Table 5: The statistics of six datasets

Dataset	# Docs	# Sentences	# References	# Comments	Abstraction	Label
DUC 01	309	10,639	60	0	Yes	Yes
DUC 02	567	15,188	116	0	Yes	Yes
DUC 04	500	13,129	200	0	Yes	Yes
TGSum (Cao et al., 2015a)	1,114	33,968	4,658	—	No	No
WG (Wei and Gao, 2014)	121	6,413	455	78,419 (tweets)	Yes	No
SoLSCSum (Nguyen et al., 2016b)	157	3,462	5,858	25,633	No	Yes
VSoLSCSum	141	3,760	2,448	6,926	No	Yes

Table 5 represents the comparison of our dataset with previous datasets in English. Compared results indicate that the number of sentences and comments in our dataset is sufficient for the summarization. In addition, our dataset includes both social information and labels annotated by human, which are not available in other datasets in Vietnamese.

2.5 Summary Generation

The summarization was formulated in the form of a learning to rank suggested by (Svore et al., 2007; Wei and Gao, 2014). To train a learning to rank model (L2R), Ranking SVM[9] (Joachims, 2006), a powerful method for information retrieval (Liu, 2011; Nguyen et al., 2016a), was adopted. Ranking SVM applies the characteristics of SVM (Cortes and Vapnik, 1995) to perform pairwise classification. Given n training queries $\{q_i\}_{i=1}^{n}$, their associated document pairs $(x_u^{(i)}, x_v^{(i)})$ and corresponding ground truth label $y_{(u,v)}^{(i)}$, Ranking SVM optimizes the objective function shown in Eq. (1):

$$\min \frac{1}{2} \|w\|^2 + \lambda \sum_{i=1}^{n} \sum_{u,v:y_{u,v}^{(i)}} \xi_{u,v}^{(i)} \tag{1}$$

$$\text{s.t. } w^T(x_u^{(i)} - x_v^{(i)}) \geqslant 1 - \xi_{u,v}^{(i)}, \text{ if } y_{u,v}^{(i)} = 1 \tag{2}$$

$$\xi_{u,v}^{(i)} \geqslant 0, i = 1, ..., n \tag{3}$$

where: $f(x) = w^T x$ is a linear scoring function, (x_u, x_v) is a pairwise and $\xi_{u,v}^{(i)}$ is the loss. The document pair-wise is sentence-sentence or comment-comment and the pair-wise order is determined by the agreement of each sentence or comment (the total label 1 over five annotators). After training, the summarization was generated by selecting top *m* ranked sentences and comments.

3 Results and Discussion

3.1 Experimental Setup

Comments with less than five tokens were eliminated since they are fairly short for summarization. 5-fold cross validation with $m = 6$ (less than 30% average sentences, see Table 3) was used.

Support Vector Machines (SVM)[10] (Cortes and Vapnik, 1995) was selected for the classification because it has shown as a competitive method for summarization. Uni-gram and bi-gram taken from KenLM[11] trained from Vietnamese data[12],[13] were used as language models for learning to rank (L2R).

3.2 Summary Systems

We validated the potential usage of our dataset on several social context summarization methods. The methods are listed as below:

- **SentenceLead:** chooses the first *x* sentences as the summarization (Nenkova, 2005).

[9] https://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html
[10] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/
[11] https://kheafield.com/code/kenlm/
[12] http://www.ted.com/talks
[13] http://vlsp.hpda.vn:8080/demo/?page=about&lang=en

- **SociVote**: selects sentences based on the voting of Cosine similarity suggested in (Wei and Gao, 2015); the threshold = 0.65.
- **LexRank:** algorithm[14] (Erkan and Radev, 2004); tokenization and stemming[15] were used.
- **cc-TAM**: built a cross-collection topic-aspect modeling (cc-TAM) as a preliminary step to generate a bipartite graph for co-ranking algorithm (Gao et al., 2012).
- **HGRW**: is a variation of LexRank named Heterogeneous Graph Random Walk (Wei and Gao, 2015); the threshold was 0.7.
- **SVM:** was used in (Yang et al., 2011; Kupiec et al., 1995; Osborne, 2002; Yeh et al., 2005). RBF kernel was used with scaling in [-1, 1].
- **RT-One Wing**: uses the features from (Nguyen and Nguyen, 2016), but only using one wing (sentences or comments) when generating the summarization. For example, when modeling a sentence, the remaining ones in the same side was utilized.
- **SoRTESum:** was proposed by (Nguyen and Nguyen, 2016) using a set of RTE similarity features (Nguyen et al., 2015). This method includes two models: SoRTESum-Inter Wing and Dual Wing.

3.3 Evaluation Metric

Gold-standard references were used for the evaluation of summary methods. Evaluation metric is F-1 of ROUGE-N[16] (N=1, 2) (Lin and Hovy, 2003).

3.4 Results and Discussion

Table 6 shows the results of summary methods on our dataset. The results indicate that: (i) our dataset benefits social context summarization in Vietnamese and (ii) social information accelerates the performance of summary methods, e.g. RTE-One Wing vs. RTE Inter Wing and Dual Wing.

Ranking SVM with local and cross features is the best in Table 6. This is because, firstly, SVMRank inherits powerful properties of SVM. For example, it can create correct margins for classification based on the help of margin maximization. In training, these properties help SVMRank to avoid an overfitting problem, which often appears in other methods, e.g. AdaBoost or RankBoost. The results of L2R using RankBoost in Table 7 support this statement. Secondly, SVMRank integrates social information leading to significant improvements compared to SVM which only uses local features, e.g. sentence length, sentence position. This also shows the efficiency of local and cross features proposed in (Wei and Gao, 2014). Finally, formulating the summarization in the form of learning to rank may be more appropriate than sentence classification, i.e. SVM.

Table 6: Summary performance on our dataset; * is supervised method; **bold** is the best value; *italic* is the second best; SentenceLead was not used in summarizing comments. Methods with *S* use social information.

System	Document						Comment					
	ROUGE-1			ROUGE-2			ROUGE-1			ROUGE-2		
	P	R	F-1	P	R	F-1	P	R	F-1	P	R	F-1
SentenceLead	0.838	0.295	0.437	0.751	0.266	0.393	—	—	—	—	—	—
SociVote (S)	0.804	0.290	0.427	0.699	0.256	0.375	0.638	0.148	0.241	0.454	0.103	0.169
LexRank	0.784	0.336	0.471	0.629	0.272	0.381	0.671	0.231	0.344	0.496	0.163	0.246
HGRW (S)	0.816	0.375	*0.514*	0.691	0.320	*0.438*	0.697	0.244	*0.362*	0.525	0.177	*0.265*
cc-TAM (S)	0.798	0.271	0.405	0.653	0.226	0.336	0.682	0.116	0.199	0.427	0.073	0.125
SVM*	0.793	0.370	0.505	0.689	0.321	*0.438*	0.511	0.237	0.324	0.309	0.127	0.181
RTE-One Wing	0.786	0.364	0.498	0.670	0.309	0.423	0.613	0.219	0.324	0.420	0.143	0.214
SoRTESum IW (S)	0.774	0.338	0.471	0.629	0.275	0.383	0.669	0.224	0.336	0.482	0.153	0.233
SoRTESum DW (S)	0.819	0.345	0.486	0.718	0.304	0.427	0.652	0.191	0.296	0.469	0.129	0.203
SVMRank* (S)	0.846	0.380	**0.525**	0.769	0.346	**0.478**	0.655	0.251	**0.364**	0.490	0.182	**0.266**

[14]https://code.google.com/p/louie-nlp/source/browse/trunk/louie-ml/src/main/java/org/louie/ml/lexrank/?r=10
[15]http://nlp.stanford.edu/software/corenlp.shtml
[16]http://kavita-ganesan.com/content/rouge-2.0-documentation

Results in Table 6 also indicate that HGRW is a competitive method, which achieves a second best result compared to Ranking SVM. This is because HGRW exploits the support of social information for the summarization. It also notes that HGRW is an unsupervised method. SVM obtains competitive results even social information was not integrated. This shows the efficiency of features for summarization in (Yang et al., 2011; Nguyen and Nguyen, 2016). SoRTESum with the support from social information obtains significant improvements as opposed to a strong method Sentence Lead, which simulates the summarization by picking up some first sentences (Nenkova, 2005). Interestingly, cc-TAM achieves the lowest result even though this method is competitive in English (Gao et al., 2012). The reason is that cc-TAM was developed for multi-document summarization but our dataset was created for single-document summarization.

3.5 The Performance of L2R Methods

The performance of Ranking SVM was compared to other L2R methods by using the same feature set (local and social features) in (Wei and Gao, 2014). The L2R methods include RankBoost (Freund et al., 2003) (*iteration* = 300, *metric* is ERR10), RankNet (Burges et al., 2005) (*epoch* = 100, *the number of layers* = 1, *the number of hidden nodes* per layer = 10 and l*earning rate* = 0.00005), Coordinate Ascent (Metzler and Croft, 2007) (random restart = 2, iteration = 25, tolerance = 0.001 with non-regularization), and Radom Forest (Breiman, 2001) (*the number of bags* = 300, *sub-sampling rate* = 1.0, *feature sampling rate* = 0.3, *ranker to bag with MART*, *the number of trees* in each bag = 100, *learning rate* = 0.1, and *the min leaf support* = 1) implemented in RankLib[17].

Table 7: The performance of L2R methods.

System	Document						Comment					
	ROUGE-1			ROUGE-2			ROUGE-1			ROUGE-2		
	P	R	F-1	P	R	F-1	P	R	F-1	P	R	F-1
RankBoost	0.820	0.366	0.507	0.717	0.324	0.447	0.663	0.242	0.355	0.498	0.175	0.259
RankNet	0.788	0.401	**0.532**	0.685	0.351	0.465	0.595	0.253	0.355	0.437	0.179	0.255
Coordinate Ascent	0.811	0.349	0.489	0.712	0.308	0.431	0.643	0.244	0.354	0.472	0.173	0.254
Random Forrest	0.847	0.374	0.520	0.771	0.343	0.475	0.649	0.252	**0.364**	0.486	0.182	0.265
SVMRank	0.846	0.380	0.525	0.769	0.346	**0.478**	0.655	0.251	**0.364**	0.490	0.182	**0.266**

Results in Table 7 illustrate that Ranking SVM (Joachims, 2006) ($C = 3$ with *linear kernel*) is the best except for ROUGE-1 in document summarization due to nice properties which Ranking SVM inherits from SVM. RankNet obtains the best result in ROUGE-1 because neural networks used in RankNet may positively affect the summarization. The remaining methods are competitive compared to results in Table 6. This concludes that formulating sentence selection as a L2R task benefits the summarization.

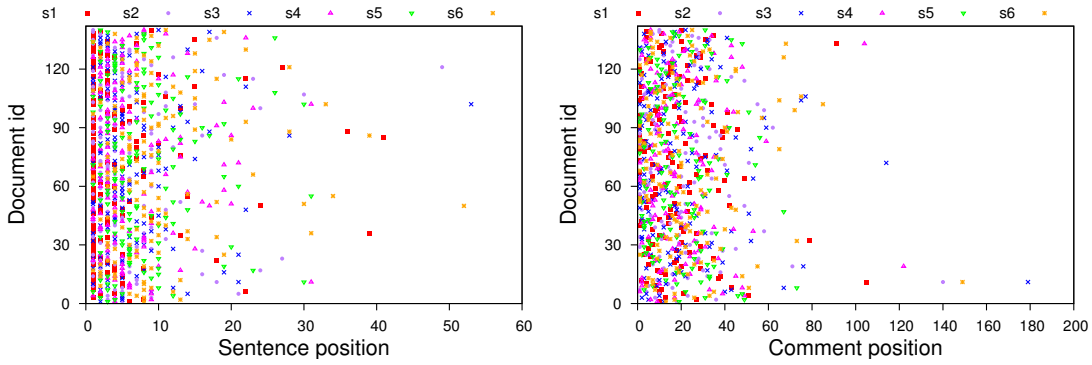
3.6 Summary Sentence Position Observation

The position of summary sentences and comments generated from Ranking SVM was observed. Figures 3a and 3b indicate that extracted sentences are within top 10 for sentences and 20 for comments. This supports the observation in Section 2.4. There also are outlier points, e.g. 52 in Figure 3a and 180 in Figure 3b. Results from Section 2.4, Figures 3a and 3b show that: (i) Sentence Lead is a competitive method (see Table 6) because Sentence Lead formulates the summarization by selecting several first sentences and (ii) this method is inefficient for comments because representative comments usually appear in a wider range in contrast to sentences. Considering Figures 2a, 2b, 3a, and 3b, we conclude that sentence position is an important feature for document summarization, not for comments.

3.7 Summary Sentence Length Observation

The average sentence length of extracted summaries generated from summary methods was also analyzed. As can be seen from Figures 4a and 4b, long sentences belong to competitive methods, e.g. SVMRank, HGRW while poor methods generate shorter sentence, e.g. cc-TAM or Sentence Lead. SVMRank obtains the longest sentences and comments, e.g. 31 in sentence and 27 in comment supporting results

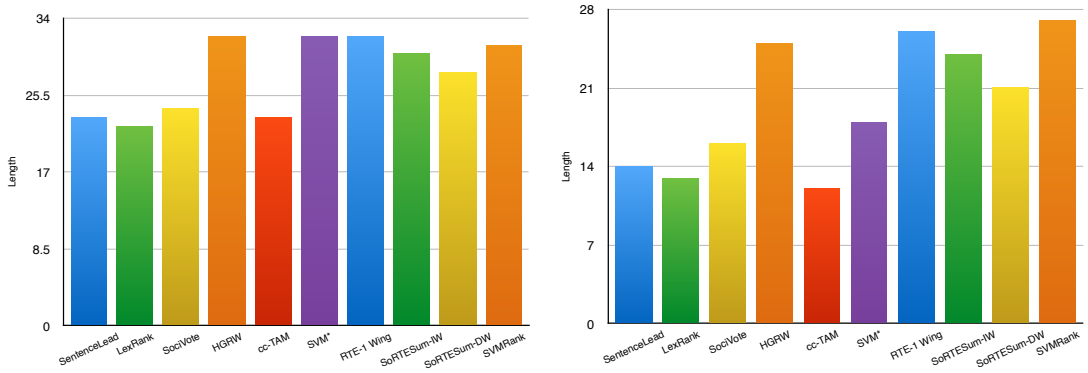
[17] http://people.cs.umass.edu/∼vdang/ranklib.html



(a) Document summarization (b) Comment summarization

Figure 3: The sentence position of extracted summaries.

in Table 6. The trend of extracted comments in Figure 3b shares the same property with sentences in Figure 3a. Considering results in Figures 4a and 4b, we conclude that sentence length is one of the most important features for the summarization.



(a) Document summarization (b) Comment summarization

Figure 4: The sentence length of extracted summaries.

4 Error Analysis

Table 8 shows the output of Ranking SVM and gold-standard references (six summary sentences and comments are shown with seven references). Ranking SVM selects three correct (denoted by [+]) and three incorrect sentences (represented by [-]) compared to the references. This is because summary sentences include vital words, e.g. *"Pokemon"*, *"driver"* and they are long sentences; as the result, local features can capture these sentences. In addition, summary sentences also share critical words with comments, e.g. *"police"*, *"Pokemon"*. In this case, cross features from comments also help to enrich information in these sentences. Nevertheless, there are several sophisticated sentences so that our model made incorrect decisions. For instance, a long non-summary sentence *S3* shares important words, e.g. *"game"* with comments.

For comment summarization, it is interesting that two comments (*C1* and *C2*) are derived from sentences. This supports the data observation in Section 2.4, which indicates that readers tend to use words or phrases appearing in article to build their comments. Since *C6* contains salient information, with local features and cross features, Ranking SVM selected this sentence correctly. Meanwhile, *C3*, *C4*, and *C5* mention readers' opinions rather than the content of the event. In this view, these comments also contribute to enrich the summarization.

Table 8: A summary example of 6^{th} Pokemon Go event document generated by Ranking SVM.

Gold-standard references	
A Vietnamese woman died on August 25 by a car accident relating to Pokemon Go game in Japan	
Daily news Mainichi announces that the victim is a 29-year-old Vietnamese woman living in Kasugai city, Aichi, Japan	
On August 11 evening, when crossing the road by bicycle, the woman was crashed by a car	
While charging his phone, he could not see the woman and the accident happened	
Playing game while driving, the driver should be suspended his driver license	
I am a game player and I really expect this game to be removed from online stores	
Addiction games are dangerous for health and money	
Summary	
Sentences	**Comments**
[+]S1: Daily news Mainichi of Japan announces that the victim is a 29-year-old Vietnamese woman living in Kasugai city, Aichi, Japan	[+]C1: The driver was released immediately after he was arrested
[+]S2: On August 11 evening, when crossing the road by her bicycle, the woman was hit by a car	[+]C2: The police is investigating the accident
[-]S3: The driver said that his phone had been out of battery due to playing the game	[-]C3: We prohibit what we cannot control, I often play this game but in a park, so there's no negative effects to other people
[+]S4: Despite driving his car, the 26-year-old driver still played Pokemon Go	[-]C4: If you don't like, you should not play because you can not give up.
[-]S5: The driver was released immediately after he was arrested	[-]C5: It depends on the responsibility of players, we can not conclude that people playing Pokemon are bad guys.
[-]S6: The police is investigating the accident	[+]C6: Driving a car while playing Pokemon, suspend their driver license rather than let get involved in a crash.

5 Conclusion

This paper presents a Vietnamese dataset named VSoLSCSum for social context summarization. The dataset is created by collecting Web documents via keywords from Vietnamese online news providers. It contains 141 documents in 12 special events. Gold-standard references are manually annotated by social users. The inter-agreement among annotators after validating calculated by Cohen's Kappa is 0.685. VSoLSCSum has two essential characteristics: (i) it includes comments as social information to support sentences for generating a high-quality summarization and (ii) it includes labels, which can be used to train supervised summary methods, e.g. SVM or L2R. Experimental results show the potential utilization of our dataset in Vietnamese social context summarization and conclude that formulating sentence selection as a L2R task benefits the summarization.

For future directions, abstractive summaries of each event should be generated. Human evaluation should also be conducted to ensure summary quality.

Acknowledgements

We would like to thank Chien-Xuan Tran for creating demo website, Gao and Li for sharing the code of (Gao et al., 2012). We also would like to thank anonymous reviewers for their detailed comments for improving our paper. This work was supported by JSPS KAKENHI Grant number 15K16048, JSPS KAKENHI Grant Number JP15K12094, and CREST, JST.

References

[Amitay and Paris2000] Einat Amitay and Ce'cile Paris. 2000. Automatically summarising web sites: is there a way around it?. In *CIKM: 173-179*.

[Breiman2001] Leo Breiman. 2001. Random forests. *Machine Learning 45(1): 5-32*.

[Burges et al.2005] Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. 2005. Learning to rank using gradient descent. In *ICML: 89-96*.

[Cao et al.2015a] Ziqiang Cao, Chengyao Chen, Wenjie Li, Sujian Li, Furu Wei, and Ming Zhou. 2015a. Tgsum: Build tweet guided multi-document summarization dataset. In *arXiv preprint arXiv:1511.08417*.

[Cao et al.2015b] Ziqiang Cao, Furu Wei, Li Dong, Sujian Li, and Ming Zhou. 2015b. Ranking with recursive neural networks and its application to multi-document summarization. In *AAAI: 2153-2159*.

[Conroy and O'Leary2001] John M. Conroy and Dianne P. O'Leary. 2001. Text summarization via hidden markov models. In *SIGIR: 406-407*.

[Cortes and Vapnik1995] Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning 20(3): 273-297*.

[Delort et al.2003] J.-Y. Delort, B. Bouchon-Meunier, and M. Rifqi. 2003. Enhanced web document summarization using hyperlinks. In *Hypertext'03: 208-215*.

[Delort2006] Jean-Yves Delort. 2006. Identifying commented passages of documents using implicit hyperlinks. In *Hypertext: 89-98*.

[Edmundson1969] Harold P. Edmundson. 1969. New methods in automatic extracting. *Journal of the Association for Computing Machinery, 16(2): 264-285*.

[Erkan and Radev2004] Gunes Erkan and Dragomir R. Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research, 22: 457-479*.

[Freund et al.2003] Yoav Freund, Raj D. Lyeryer, Robert E. Schapire, and Yoram Singer. 2003. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research 4: 933-969*.

[Ganesan et al.2010] Kavita Ganesan, ChengXiang Zhai, and Jiawei Han. 2010. Opinosis: a graph-based approach to abstractive summarization of highly redundant opinions. In *COLLING: 340-348*.

[Gao et al.2012] Wei Gao, Peng Li, and Kareem Darwish. 2012. Joint topic modeling for event summarization across news and social media streams. In *CIKM:1173-1182*.

[Hu et al.2007] Meishan Hu, Aixin Sun, and Ee-Peng Lim. 2007. Comments-oriented blog summarization by sentence extraction. In *CIKM: 901-904*.

[Hu et al.2008] Meishan Hu, Aixin Sun, and Ee-Peng Lim. 2008. Comments-oriented document summarization: Understanding document with readers' feedback. In *SIGIR: 291-298*.

[Joachims2006] Thorsten Joachims. 2006. Training linear svms in linear time. In *KDD: 217-226*.

[Kim and Zhai2009] Hyun Duk Kim and ChengXiang Zhai. 2009. Generating comparative summaries of contradictory opinions in text. In *CIKM: 385-394*.

[Kupiec et al.1995] Julian Kupiec, Jan O. Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *SIGIR: 68-73*.

[Lafferty et al.2001] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML: 282-289*.

[Lin and Hovy2003] Chin-Yew Lin and Eduard H. Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *HLT-NAACL: 71-78*.

[Liu2011] Tie-Yan Liu. 2011. *Learning to Rank for Information Retrieval*. Springer, ISBN 978-3-642-14266-6, pp. I-XVII, 1-285.

[Lu et al.2009] Yue Lu, ChengXiang Zhai, and Neel Sundaresan. 2009. Rated aspect summarization of short comments. In *WWW: 131-140*.

[Luhn1958] Hans P. Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research Development, 2(2): 159-165*.

[Metzler and Croft2007] Donald Metzler and W. Bruce Croft. 2007. Linear feature-based models for information retrieval. *Inf. Retr. 10(3): 257-274*.

[Nenkova2005] Ani Nenkova. 2005. Automatic text summarization of newswire: lessons learned from the document understanding conference. In *AAAI: 1436-1441*.

[Nguyen and Nguyen2016] Minh-Tien Nguyen and Minh-Le Nguyen. 2016. Sortesum: A social context framework for single-document summarization. In *ECIR: 3-14*.

[Nguyen et al.2015] Minh-Tien Nguyen, Quang-Thuy Ha, Thi-Dung Nguyen, Tri-Thanh Nguyen, and Le-Minh Nguyen. 2015. Recognizing textual entailment in vietnamese text: An experimental study. In *KSE: 108-113*.

[Nguyen et al.2016a] Minh-Tien Nguyen, Viet-Anh Phan, Truong-Son Nguyen, and Minh-Le Nguyen. 2016a. Learning to rank questions for community question answering with ranking svm. In *CoRR abs/1608.04185*.

[Nguyen et al.2016b] Minh-Tien Nguyen, Chien-Xuan Tran, Duc-Vu Tran, and Minh-Le Nguyen. 2016b. Solscsum: A linked sentence-comment dataset for social context summarization. In *CIKM: 2409-2412*.

[Nguyen et al.2016c] Minh-Tien Nguyen, Duc-Vu Tran, Chien-Xuan Tran, and Minh-Le Nguyen. 2016c. Learning to summarize web documents using social information. In *ICTAI*.

[Osborne2002] Miles Osborne. 2002. Using maximum entropy for sentence extraction. In *ACL Workshop on Automatic Summarization: 1-8*.

[Paul et al.2010] Michael J. Paul, ChengXiang Zhai, and Roxana Girju. 2010. Summarizing contrastive viewpoints in opinionated text. In *EMNLP: 66-76*.

[Shen et al.2007] Dou Shen, Jian-Tao Sun, Hua Li, Qiang Yang, and Zheng Chen. 2007. Document summarization using conditional random fields. In *IJCAI: 2862-2867*.

[Sun et al.2005] Jian-Tao Sun, Dou Shen, Hua-Jun Zeng, Qiang Yang, Yuchang Lu, and Zheng Chen. 2005. Webpage summarization using clickthrough data. In *SIGIR: 194-201*.

[Svore et al.2007] Krysta Marie Svore, Lucy Vanderwende, and Christopher J. C. Burges. 2007. Enhancing single-document summarization by combining ranknet and third-party sources. In *EMNLP-CoNLL: 448-457*.

[Wei and Gao2014] Zhongyu Wei and Wei Gao. 2014. Utilizing microblogs for automatic news highlights extraction. In *COLING: 872-883*.

[Wei and Gao2015] Zhongyu Wei and Wei Gao. 2015. Gibberish, assistant, or master?: Using tweets linking to news for extractive single-document summarization. In *SIGIR: 1003-1006*.

[Yang et al.2011] Zi Yang, Keke Cai, Jie Tang, Li Zhang, Zhong Su, and Juanzi Li. 2011. Social context summarization. In *SIGIR: 255-264*.

[Yeh et al.2005] Jen-Yuan Yeh, Hao-Ren Ke, Wei-Pang Yang, and I-Heng Meng. 2005. Text summarization using a trainable summarizer and latent semantic analysis. *Inf. Process. Manage. 41(1): 75-95*.