Using Word Sequences for Text Summarization

Esaú Villatoro-Tello, Luis Villaseñor-Pineda, and Manuel Montes-y-Gómez

Language Technologies Group, Computer Science Department,
National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico
{villatoroe, villasen, mmontesg}@inaoep.mx

Abstract. Traditional approaches for extractive summarization score/classify sentences based on features such as position in the text, word frequency and cue phrases. These features tend to produce satisfactory summaries, but have the inconvenience of being domain dependent. In this paper, we propose to tackle this problem representing the sentences by word sequences (n-grams), a widely used representation in text categorization. The experiments demonstrated that this simple representation not only diminishes the domain and language dependency but also enhances the summarization performance.

1 Introduction

Current information technologies allow the creation and storage of massive amounts of data. In this context, document summaries are becoming essential. People can explore and analyze entire document collections just by looking at their summaries [1].

Text summarization is the task concerning the automatic generation of document summaries. It aims to reduce documents in length and complexity, while preserving some of their essential information [2]. Despite there are different types of summaries and approaches for their generation, today the most popular summarization systems focus on the construction of extractive summaries (extracts created by selecting a set of relevant sentences of the input text) by machine-learning techniques [3].

One central problem in machine-learning summarization is the representation of sentences. There have been used several surface-level features in order to represent them. Most of these features are "heuristically motivated", since they tend to emulate the manual creation of extracts. In a pioneering work by Kupiec *et al.* [2] sentences were represented by their position and length, the presence of cue phrases and their overlap with the document title. More recent works [1,4] enlarged these features incorporating information such as the occurrence of proper names and the presence of anaphors.

The "heuristically motivated" features allow producing very precise extracts. Nevertheless, they have the major disadvantage of being highly related to a target domain. This condition implies that when moving from one domain to another, it may be necessary to redefine or even eliminate some features. For instance, cue phrases, which are particular for each domain, require being modified, while the overlap with the title, which has no sense in all topics, may be eliminated.

In order to increase the domain (and language) independence of machine learning summarizers, we propose eliminating all kind of "heuristically motivated" attributes and substitute them by word-based features. In particular, we consider the use of word sequences

Petr Sojka, Ivan Kopeček and Karel Pala (Eds.): TSD 2006, LNAI 4188, pp. 293–300, 2006.

(so-called *n*-grams) as sentence features. Our goal is to develop a more flexible and competitive summarization method. In other words, we aim to boost the summarization flexibility without reducing the quality of the output summaries.

It is important to mention that simple word-based representations are common in many text-processing tasks. However, *n*-grams have been applied without much significant success. In this way, one relevant contribution of this work is the study of the application of word-based representations in text summarization, and the evaluation of the impact of using word sequences as sentence features. In our knowledge, this is the first attempt on using word sequence features for broad-spectrum text summarization.

The rest of the paper is organized as follows. Section 2 introduces the proposed feature scheme. Section 3 describes the experimental setup. Section 4 presents some experimental results on the use of word sequences as features for text summarization. Finally, section 5 depicts our conclusions and future work.

2 Word-Based Features

As we mentioned, the machine-learning approach for text summarization focuses on the creation of extracts by the selection of relevant sentences from the input texts. To pursue this approach it is necessary to establish the sentence features, the classification method and a training corpus of document/extract pairs.

Traditional methods for supervised text summarization use "heuristically motivated" features to represent the sentences. Our proposal is to consider word-based features in order to increase the summarization flexibility by lessening the domain and language dependency. In particular, we propose using *n*-grams (sequences of *n* consecutive words) as sentence features. Thus, in our model each sentence is represented by a feature vector that contains one boolean attribute for each *n*-gram that occurs in the training collection. Specially, we only consider sequences up to three words, i.e., from *1*-grams to *3*-grams.

Word-based representations have been widely used in several text-processing tasks. In particular, in text categorization the bag-of-words (*1*-grams) representation corresponds to the leading approach [5]. However, there are numerous studies on the effect of generalizing this approach by using word sequences as document features [6,7,8]. These studies indicate that the use of word *n*-grams does not considerably improve the performance on text categorization.

Despite of the unfavorable results in text categorization, we believe that the use of *n*-grams can be helpful in text summarization. This hypothesis is supported in the following two facts:

On the one hand, sentences are much smaller than documents, and consequently the classifier would require more and more detailed information to distinguish between relevant and irrelevant instances. For instance, in text categorization, the merely presence of the word earthquake may indicate that the document at hand is about this phenomenon. Nevertheless, it may not be enough to select the informative sentences. In text summarization, *n*-grams such as "earthquake-left" or "earthquake-of-magnitude" are more pertinent.

On the other hand, some recent works on text summarization make use of *n*-grams to evaluate the quality of summaries [9,10]. These works have shown that the *n*-gram correspondences between handwritten and automatically produced summaries are a good indicator of the appropriateness of the extracts.

Our proposal differs from these works in that it directly employs the n-grams to construct the summaries, i.e., it uses the n-grams to select the relevant sentences. Therefore, it represents the first attempt on using word sequence features in text summarization, and consequently the first evaluation on their impact in the quality of the extracts.

3 Experimental Setup

3.1 Corpora

We used two different corpora in our experiments, one of them in Spanish and the other in English. Both corpora consist of newspaper articles, but the first one only includes news about natural disasters, while the other considers different kinds of topics such as politics, economics and sports. Table 1 resumes some statistics about the corpora.

Table 1. Corpora Statistics

Data Set	Language	Domain	Number of Sentences	Relevant Sentences
DISASTERS	Spanish	Natural Disasters News	2833	863 (30%)
CAST	English	General News	4873	1316 (27%)

The *Disasters* corpus consists of 300 news reports collected from several Mexican newspapers. Each sentence of the corpus was labeled using two basic tags: relevant and non-relevant. In order to avoid subjectivity on the tagging process, annotators were instructed to mark as relevant only the sentences containing at least one concrete fact about the event. For instance, the date or place of the disaster occurrence, or the number of people or houses affected.

On the other hand, the *CAST* (Computer-Aided Summarization) corpus consists of 164 news reports. In contrast to the Disasters corpus, it includes news about different topics such as politics, economics and sports. It sentences were also annotated as relevant and non-relevant. Both corpora maintain a similar distribution of relevant sentences. More details on the *CAST* corpus can be found in [11].

3.2 Classifier

The Naïve Bayes classifier has proved to be quite competitive for most text processing tasks including text summarization. This fact supported our decision to use it as main classifier for our experiments. It basically computes for each sentence s its probability (i.e., a score) of been included in a summary S given the k features $F_j; j = 1..k$. This probability can be expressed using Bayes' rule as follows [2]:

$$P(s \in S|F_1, F_2, ..., F_k) = \frac{P(F_1, F_2, ..., F_k|s \in S)P(s \in S)}{P(F_1, F_2, ..., F_k)}$$

Assuming statistical independence of the features:

$$P(s \in S|F_1, F_2, ..., F_k) = \frac{\prod_{j=1}^{k} P(F_j|s \in S)P(s \in S)}{\prod_{j=1}^{k} P(F_j)}$$

where $P(s \in S)$ is a constant and $P(F_j|s \in S)$ and $P(F_j)$ can be estimated directly from the training set by counting occurrences.

3.3 Baseline Configuration

In order to define the baseline configuration we made an exhaustive study of previous supervised methods for text summarization. Particularly, we searched for common features across the different methods as well as for *domain independent* features. The following paragraphs briefly describe our main findings.

Kupiec *et al.* [2] used five different attributes, but only three of them were domain independent, namely, the position and length of the sentence, and the presence of proper names.

Chuang *et al.* [1] evaluated the representation of sentences by 23 different features. However, only a small subset of them were domain independent. For instance, it used the similarity with the document title and the term frequencies.

Neto *et al.* [4] used 13 features in their summarization system. Only four of them were domain independent: the centroid value of the sentence, its length and position, as well as the similarity with the title and the presence of proper names.

We implemented a baseline summarization method using the following features: the position and length of sentences, its centroid value and its similarity with the document title, and the presence of proper names. All these features are domain and language independent, and thus they may be applied to both corpora.

In addition, we also included the presence of numeric quantities. This feature was added because both data sets are news articles and they tend to use numeric expressions to explain the facts.

4 Experimental Results

In this paper, we have proposed the use of word-based features in order to develop a more flexible and competitive summarization method. This section presents the results of two initial experiments. The first experiment considers the representation of sentences by simple bag-of-words. Its purpose is to demonstrate that word-based features are domain and language independent and that its performance is comparable to that of traditional approaches. The second experiment applies word sequences as sentence features. Its goal is to evaluate their impact on text summarization.

In both experiments, the performance of classifiers was measured by the accuracy, precision and recall, and the evaluation was based on a cross-validation strategy.

4.1 First Experiment: Single Words as Features

In this experiment, single-word features represented sentences. Since the original feature space had a very high dimensionality, we needed to apply the information gain technique in order to select a subset of relevant features. Table 2 shows the number of features considered in this experiment for both data sets.

Table 3 presents the results obtained in this experiment. It is important to notice that (*i*) the proposed representation produced a similar performance for both data sets, indicating that it

Table 2. Number of single-word features

	Original Features	**Selected Features**
DISASTERS	8958	530
CAST	10410	612

Table 3. Evaluation of single-word features

	Baseline Configuration			**Single-Word Features**		
	accuracy	*precision*	*recall*	*accuracy*	*precision*	*recall*
DISASTERS	74.94	87.89	78.89	84.82	91.72	87.12
CAST	68.08	74.36	80.44	79.76	88.67	84.39

is domain and language independent, and that (*ii*) the proposed representation outperformed the baseline method, in both precision and recall.

4.2 Second Experiment: Word Sequences as Features

Here, we represented sentences by word sequences (n-grams). Specifically, we considered sequences up to three words, i.e., from *1*-grams to *3*-grams. Like in the previous experiment, we used the information gain technique to reduce the feature space and to select a subset of relevant features. Table 4 shows the number of features considered in this experiment for both data sets.

Table 4. Number of word sequence features

	Original Features				**Selected Features**
	1-grams	*2-grams*	*3-grams*	*All*	*All*
DISASTERS	8958	34340	53356	96654	2284
CAST	10410	52745	72953	136108	2316

Table 5 describes the results obtained in this experiment. They indicate that the use of n-gram features enhanced the classification precision, while maintaining the recall rate. This behavior is a direct cause of using features that are more detailed. This kind of features allows a better distinction between relevant and non-relevant sentences. In particular, they allow treating difficult cases.

Table 5. Evaluation of word sequence features

	Single-Word Features			**Word sequence features**		
	accuracy	*precision*	*recall*	*accuracy*	*precision*	*recall*
DISASTERS	84.82	91.72	87.12	86.16	95.53	86.09
CAST	79.76	88.67	84.39	84.54	96.48	84.53

4.3 A Practical Example

This section illustrates the summarization based on word sequence features. In particular, table 6 shows a news article from the CAST corpus and it corresponding calculated extract (in bold font).

Table 6. A document and it corresponding extract

Sentence ID	Relevance Assessments	Sentences
1	×	USA: U.S. June trade gap narrows sharply as imports drop.
2	×	U.S. June trade gap narrows sharply as imports drop.
3	×	Glenn Somerville.
4	×	WASHINGTON 1996-08-20.
5	✓	**The U.S. trade gap narrowed dramatically in June as imports of merchandise and petroleum plunged from May levels, the Commerce Department said on Tuesday.**
6	✓	**The monthly deficit dropped 23.1 percent to $8.11 billion from a revised $10.55 billion in May much lower than the $9.4 billion shortfall that Wall Street economists had forecast for June.**
7	✓	June exports eased a slight 0.3 percent to $69.71 billion while imports dropped 3.3 percent to $77.82 billion.
8	✓	**Amid the big overall improvement in June trade, China emerged for the first time as the nation with which the United States has the largest bilateral shortfall.**
9	×	The deficit with China climbed 8.8 percent to $3.33 billion in June, surpassing the $3.24 billion deficit with Japan that was up 3.6 percent from May.
10	×	**Commerce noted that exports of American-made goods to China declined for a fourth straight month in June, which is likely to fuel trade tensions between the two countries.**
11	✓	**Steady improvement in shrinking the deficit with Japan was the main reason that China became the leading deficit nation in June, Commerce officials said.**
12	×	The second-quarter deficit of $10.5 billion with Japan was the smallest quarterly deficit in five years, the department said.
13	×	Previously, the department said the overall May trade deficit was $10.88 billion but it revised that down to a $10.55 billion gap.
14	✓	**The United States typically runs a surplus on its trade with other countries in services like travel and tourism that partly offsets big merchandise trade deficits.**
15	✓	In June, the merchandise deficit fell 13.9 percent to $14.46 billion from $16.79 billion in May.
16	×	Lower imports of new cars and parts, especially from Japan and Germany, helped shrink the merchandise trade gap.
17	×	The surplus on services climbed 1.6 percent to $6.34 billion from $6.25 billion in May.
18	×	Analysts said beforehand that an influx of tourists bound for the Olympic Games in Atlanta would boost the services surplus.
19	×	The cost and volume of all types of petroleum products fell in June after a sharp May runup.
20	×	The cost of petroleum imports declined to $5.33 billion in June from $5.93 billion while the volume fell to 291,866 barrels from 305,171 in May.
21	×	Foreign sales of civilian aircraft declined in June by $117 million to $1.54 billion.
22	×	Exports of industrial supplies and materials were off 138$million to$12.32 billion.
23	×	Imports of autos and parts from all sources dropped sharply by $689 million to $10.79 billion in June.
24	×	Computer imports were down $413 million to $4.24 billion and semiconductor imports decreased $291 million to $2.87 billion in June.
25	×	In bilateral trade, the deficit with Western Europe fell 7.1 percent to $761 million and the shortfall with Canada was down 2.2 percent to $2.42 billion.
26	×	In trade with Mexico, the U.S. deficit shrank 6.4 percent to $1.49 billion amid signs the Mexican economy was recovering from a deep recession and grew solidly in the second quarter this year.
27	×	The deficit with oil-producing OPEC countries dropped 26.9 percent in June to $1.40 billion from $1.91 billion in May.

It is important to notice that each sentence of the article has associated a manual relevance judgment (✓ for relevant sentences and × for non-relevant ones), and that the summarization procedure could identify most of the relevant sentences and just misclassified three sentences (7, 10 and 15). The generated extract contains six sentences, achieving a compression rate of 22%, and a precision and recall of 0.83 and 0.71 respectively.

5 Conclusions

This paper proposed the use of word-based features in text summarization. Specifically, it considered the use of word sequence (n-gram) features. Its goal was to increase the domain (and language) independence of machine-learning summarizers, and to develop a more flexible and competitive summarization method.

The main contributions of this paper were the following two:

On the one hand, it represented, in our knowledge, the first attempt on using word-based features for broad-spectrum text summarization. In this line, our conclusion was that these features are as appropriate for text summarization as they are for text categorization. In our experiments, they outperformed the baseline method, in both precision and recall. In addition, they were appropriated for both domains and both languages.

On the other hand, this paper presented an evaluation of the impact of using word sequences (n-grams) as sentence features in text summarization. In contrast to text categorization, where the application of n-grams has not improved the classification performance, our results confirmed that the n-grams are helpful in text summarization. In particular, these results indicated that the n-gram features enhanced the classification precision, while maintaining the recall rate. Our general conclusion in this line is that n-gram features are adequate for fine-grained classification tasks.

Acknowledgements. This work was done under partial support of CONACYT (scholarship 189943, project grants 43990 and U39957-Y). We also thank SNI-Mexico and INAOE for their assistance.

References

1. Chuang T. W., and Yang J. (2004). Text Summarization by Sentence Segment Extraction Using Machine Learning Algorithms. In *Proceedings of the ACL'04 Workshop.* Barcelona, España, 2004.
2. Kupiec, J., Pedersen J. O., and Chen, F. (1995). A Trainable Document Summarizer. In *Proceedings of the 18th ACM-SIGIR Conference on Research and Development in Information Retrieval.* Seattle, pp. 68–73, 1995.
3. Hovy, E. (2003) Text Summarization. In Mitkov R. (Ed). *The Oxford handbook of Computational Linguistics.* Oxford, NY, 2003.
4. Neto L., Freitas A. A., and Kaestner C. A. A. (2004). Automatic Text Summarization using a Machine Learning Approach. In *Proceedings of the ACL-04 Workshop.* Barcelona, España, 2004.
5. Sebastiani F. (1999) Machine Learning in Automated Text Categorization. In *ACM Computing Surveys*, Vol. 34, pp. 1–47, 1999.
6. Bekkerman R., and Allan J. (2003). Using Bigrams in Text Categorization. *Technical Report IR-408.* Departement of Computer Science, University of Masschusetts, USA, 2003.

7. Canvar W. B., and Trenkle J. M. (1994). N-Gram-Based Text Categorization. In *Proceedings of the third Annual Symposium on Document Analysis and Information retrieval.* Nevada, Las Vegas, pp. 161–169, 1994.
8. Fürnkranz J. (1998). A Study Using $n-$gram Features for Text Categorization. *Technical report OEFAI-TR-98-30.* Austrian Institute for Artificial Intelligence, Wien, Austria, 1998.
9. Lin C., and Hovy E. (2003). Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of the Human Technology Conference 2003.* Edmonton, Canada, 2003.
10. Banko M., and Vanderwende L. (2004). Using N-grams to Understand the Nature of Summaries. In *Proceedings of HLT/NAACL 2004*. Boston, MA., 2004.
11. Hasler L., Orasan C. and Mitkov R. (2003): Building better corpora for summarisation. In *Proceedings of Corpus Linguistics 2003*, Lancaster, UK, pp. 309–319, 2003.