Automatic Evaluation of Summaries Using Document Graphs

Eugene Santos Jr., Ahmed A. Mohamed, and Qunhua Zhao
Computer Science and Engineering Department
University of Connecticut
191 Auditorium Road, U-155, Storrs, CT 06269-3155
{eugene, amohamed, qzhao}@engr.uconn.edu

Abstract

Summarization evaluation has been always a challenge to researchers in the document summarization field. Usually, human involvement is necessary to evaluate the quality of a summary. Here we present a new method for automatic evaluation of text summaries by using document graphs. Data from Document Understanding Conference 2002 (DUC-2002) has been used in the experiment. We propose measuring the similarity between two summaries or between a summary and a document based on the concepts/entities and relations between them in the text.

1 Introduction

Document summarization has been the focus of many researchers for the last decade, due to the increase in on-line information and the need to find the most important information in a (set of) document(s). One of the biggest challenges in text summarization research is how to evaluate the quality of a summary or the performance of a summarization tool. There are different approaches to evaluate overall quality of a summarization system. In general, there are two types of evaluation categories: *intrinsic* and *extrinsic* (Sparck-Jones and Galliers, 1996). Extrinsic approaches measure the quality of a summary based on how it affects certain tasks. In intrinsic approaches, the quality of the summarization is evaluated based on analysis of the content of a summary itself. In both categories human involvement is used to judge the summarization outputs. The problem with having humans involved in evaluating summaries is that we can not hire human judges every time we want to evaluate summaries (Mani and Maybury, 1999). In this paper, we discuss a new automated way to evaluate machine-generated summaries without the need to have human judges being involved which decreases the cost of determining which summarization system is best. In our experiment, we used data from Document Understanding Conference 2002 (DUC-2002).

2 Related Work

Researchers in the field of document summarization have been trying for many years to define a metric for evaluating the quality of a machine-generated summary. Most of these attempts involve human interference, which make the process of evaluation expensive and time-consuming. We discuss some important work in the intrinsic category.

2.1 Sentence Precision-Recall Measure

Sentence precision and recall have been widely used to evaluate the quality of a summarizer (Jing et al., 1998). Sentence precision measures the percent of the summary that contains sentences matched with the model summary. Recall, on the other hand, measures the percent of sentences in the ideal summary that have been recalled in the summary. Even though sentence precision/recall factors can give us an idea about a summary's quality, they are not the best metrics to evaluate a system's quality. This is due to the fact that a small change in the output summary can dramatically affect the quality of a summary (Jing et al., 1998). For example, it is possible that a system will pick a sentence that does not match with a model sentence chosen by an assessor, but is

equivalent to it in meaning. This, of course, will affect the score assigned to the system dramatically. It is also obvious that sentence precision/recall is only applicable to the summaries that are generated by sentence extraction, not abstraction (Mani, 2001).

2.2 Content-Based Measure

Content-based measure computes the similarity at the vocabulary level (Donaway, 2000 and Mani, 2001). The evaluation is done by creating term frequency vectors for both the summary and the model summary, and measuring the cosine similarity (Salton, 1988) between these two vectors. Of course, the higher the cosine similarity measure, the higher the quality of the summary is. Lin and Hovy (2002) used accumulative n-gram matching scores between model summaries and the summaries to be evaluated as a performance indicator in multi-document summaries. They achieved their best results by giving more credit to longer n-gram matches with the use of Porter stemmer.

A problem raised in the evaluation approaches that use the cosine measure is that the summaries may use different key terms than those in the original documents or model summaries. Since term frequency is the base to score summaries, it is possible that a high quality summary will get a lower score if the terms used in the summary are not the same terms used in most of the document's text. Donaway et al. (2000) discussed using a common tool in information retrieval: latent semantic indexing (LSI) (Deerwester et al., 1990) to address this problem. The use of LSI reduces the effect of near-synonymy problem on the similarity score. This is done by penalizing the summary less in the reduced dimension model when there are infrequent terms synonymous to frequent terms. LSI averages the weights of terms that co-occur frequently with other mutual terms. For example, both "bank" and "financial institution" often occur with the term "account" (Deerwester et al., 1990). Even though using LSI can be useful in some cases, it can produce unexpected results when the document contains terms that are not synonymous to each other, but, however, they co-occur with other mutual terms.

2.3 Document Graph

2.3.1 Representing Content by Document Graph

Current approaches in content-based summarization evaluation ignore the relations between the keywords that are expressed in the document. Here, we introduce our approach, which measures the similarity between two summaries or a summary and a document based on the relations (between the keywords). In our approach, each document/summary is represented as a document graph (DG), which is a directed graph of concepts/entities and the relations between them. A DG contains two kinds of nodes, concept/entity nodes and relation nodes. Currently, only two kinds of relations, "*isa*" and "*related to*", are captured (Santos et al, 2001) for simplicity.

To generate a DG, a document/summary in plain text format is first tokenized into sentences; and then, each sentence is parsed using Link Parser (Sleator and Temperley, 1993), and the noun phrases (NP) are extracted from the parsing results. The relations are generated based on three heuristic rules:

- The NP-heuristic helps to set up the hierarchical relations. For example, from a noun phrase "folk hero stature", we generate relations "folk hero stature *isa* stature", "folk hero stature *related to* folk hero", and "folk hero *isa* hero".
- The NP-PP-heuristic attaches all prepositional phrases to adjacent noun phrases. For example, from "workers at a coal mine", we generate a relation, "worker *related to* coal mine".
- The sentence-heuristic relates concepts/entities contained in one sentence. The relations created by sentence-heuristic are then sensitive to verbs, since the interval between two noun phrases usually contains a verb. For example, from a sentence "Workers at a coal mine went on strike", we generate a relation "worker *related to* strike". Another example, from "The usual cause of heart attacks is a blockage of the coronary arteries", we generate "heart attack cause *related to* coronary artery blockage". Figure 1 shows a example of a partial DG.

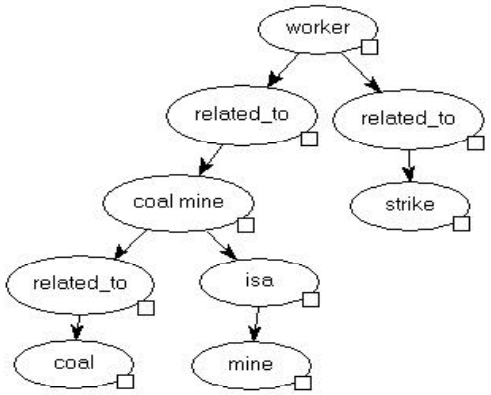

Figure 1: A partial DG.

2.3.2 Similarity Comparison between two Document Graphs

The similarity of DG_1 to DG_2 is given by the equation:

$$Sim(DG_1, DG_2) = \frac{n}{2N} + \frac{m}{2M}$$

which is modified from Montes-y-Gómez et al. (2000). N is the number of concept/entity nodes in DG_1, and M stands for number of relations in DG_1; n is the number of matched concept/entity nodes in two DGs, and m is the number of matched relations. We say we find a matched relation in two different DGs, only when both of the two concept/entity nodes linked to the relation node are matched, and the relation node is also matched. Since we might compare two DGs that are significantly different in size (for example, DGs for an extract vs. its source document), we used the number of concept/entity nodes and relation nodes in the target DG as N and M, instead of the total number of nodes in both DGs. The target DG is the one for the extract in comparing an extract with its source text. Otherwise, the similarity will always be very low. Currently, we weight all the concepts/entities and relations equally. This can be fine tuned in the future.

3 Data, and Experimental Design

3.1 Data

Because the data from DUC-2003 were short (~100 words per extract for multi-document task), we chose to use multi-document extracts from DUC-2002 (~200 words and ~400 words per extract for multi-document task) in our experiment. In this corpus, each of ten information analysts from the National Institute of Standards and Technology (NIST) chose one set of newswire/paper articles in the following topics (Over and Liggett, 2002):

- A single natural disaster event with documents created within at most a 7-day window
- A single event of any type with documents created within at most a 7-day window
- Multiple distinct events of the same type (no time limit)
- Biographical (discuss a single person)

Each assessor chose 2 more sets of articles so that we ended up with a total of 15 document sets of each type. Each set contains about 10 documents. All documents in a set are mainly about a specific "concept."

A total of ten automatic summarizers participated to produce machine-generated summaries. Two extracts of different lengths, 200 and 400 words, have been generated for each document-set.

3.2 Experimental Design

A total of 10 different automatic summarization systems submitted their summaries to DUC. We obtained a ranking order of these 10 systems based on sentence precision/recall by comparing the machine generated extracts to the human generated model summaries. The F-factor is calculated from the following equation (Rijsbergen, 1979):

$$F = \frac{2 \times P \times R}{(P + R)}$$

where P is the precision and R is the recall. We think this ranking order gives us some idea on how human judges think about the performance of different systems.

For our evaluation based on DGs, we also calculated F-factors based on precision and recall, where $P = Sim(DG_1, DG_2)$ and $R = Sim(DG_2, DG_1)$. In the first experiment, we ranked the 10 automatic summarization systems by comparing DGs generated from their outputs to the DGs generated from model summaries. In this case, DG_1 is the machine generated extract and DG_2 is the human generated extract. In the second experiment, we ranked the systems by comparing machine generated extracts to the original documents. In this case, DG_1 is an extract and DG_2 is the corresponding original document. Since the extracts were generated from multi-document sets, we

System rank	Sentence-based Ranking	Sentence-based *F*-factor	DG-based Ranking	DG-based F-factor
1 (worst)	22	0.000	22	0.122
2	16	0.062	16	0.167
3	31	0.081	31	0.180
4	25	0.081	25	0.188
5	29	0.090	29	0.200
6	20	0.125	20	0.226
7	28	0.138	28	0.255
8	24	0.171	19	0.283
9	19	0.184	24	0.283
10 (best)	21	0.188	21	0.308

Table 1: Model Summaries vs. machine-generated summaries. Ranking results for 200 words extracts

System rank	Sentence-based Ranking	Sentence-based *F*-factor	DG-based Ranking	DG-based F-factor
1 (worst)	22	0.000	22	0.181
2	16	0.128	16	0.235
3	25	0.147	25	0.256
4	31	0.150	31	0.266
5	29	0.155	20	0.273
6	20	0.172	29	0.279
7	28	0.197	28	0.316
8	24	0.223	19	0.337
9	19	0.224	24	0.355
10 (best)	21	0.258	21	0.372

Table 2: Model Summaries vs. machine-generated summaries. Ranking results for 400 words extracts

used the average of the *F*-factors for ranking purposes.

4 Results

The ranking orders obtained based on sentence precisions and recalls are shown in Tables 1 and 2. The results indicate that for sentence precision and recall, the ranking order for different summarization systems is not affected by the summarization compression ratio. The ranking results for 200-word extracts and 400-word extracts are exactly the same.

Since the comparison is between the machine generated extracts and the human created model extracts, we believe that the rankings should represent the performance of 10 different automated summarization systems, to some degree. The experiments using DGs instead of sentence matching give two very similar ranking orders (Spearman rank correlation coefficient [Myers and Well, 1995] is 0.988) where only systems 24 and 19 are reversed in their ranks (Tables 1 and 2). The results show that when the evaluation is based on the comparison between machine generated extracts and the model extracts, our DG-based evaluation approach will provide roughly the same ranking results as the sentence precision and recall approach. Notice that the *F*-factors obtained by experiments using DGs are higher than those calculated based on sentence matching. This is because our DG-based evaluation approach compares the two extracts at a more fine grained level than sentence matching does since we compare the similarity at the level of concepts/entities and their relations, not just whole sentences. The similarity of the two extracts should actually be higher than the score obtained with sentence matching because there are sentences that are equivalent in meaning but not syntactically identical.

Since we believe that the DGs captures the semantic information content contained in the respective documents, we rank the automatic summarization systems by comparing the DGs of their extract outputs against the DGs of the original documents. This approach does not need the model summaries, and hence no human involvement is needed in the evaluation. The results are shown in Tables 3 and 4. As we can see, our rankings are different from the ranking results based on comparison against the model extracts. System 28 has the largest change in rank in both 200-word and 400-word summaries. It was ranked as the worst by our DG based approach instead of number 7 (10 is the best) by the approaches comparing to the model extracts. We investigated the extract content of system 28 and found that many extracts

generated by system 28 included sentences that contain little information, e.g., author's names, publishers, date of publication, etc. The following are sample extracts produced for document 120 by systems 28, 29 (the best ranked) and a human judge, at 200-words.

[Extract for Document 120 by System 28]
John Major, endorsed by Margaret Thatcher as the politician closest to her heart, was elected by the Conservative Party Tuesday night to succeed her as prime minister.
Hong Kong WEN WEI PO
By MICHAEL CASSELL and IVOR OWEN
By MICHAEL THOMPSON-NOEL
By DOMINIC LAWSON
From Times Wire Services
By WILLIAM TUOHY, TIMES STAFF WRITER
From Associated Press

[Extract for Document 120 by System 29]
John Major, endorsed by Margaret Thatcher as the politician closest to her heart, was elected by the Conservative Party Tuesday night to succeed her as prime minister.
Aides said Thatcher is "thrilled".
Hurd also quickly conceded.
ONE year ago tomorrow, Mr John Major surprised everyone but himself by winning the general election.
It has even been suggested that the recording of the prime minister's conversation with Michael Brunson, ITN's political editor, in which Major used a variety of four-, six- and eight-letter words to communicate his lack of fondness for certain colleagues, may do him good.
BFN
[Colin Brown article: "Cabinet Allies Close Ranks But Bring
Right-wing MPs confirmed the findings in an INDEPENDENT ON SUNDAY/NOP [National Opinion Poll] poll that Michael Heseltine was the favourite to replace Mr Major, if he is forced out.
The Labour Party controls 90 local councils, whereas the Conservatives only control 13, with a sharp contrast in strength between the two sides.
If he did not see the similarity, that is still more revealing.

[Extract for Document 120 by a human judge -- model extract]
John Major, endorsed by Margaret Thatcher as the politician closest to her heart, was elected by the Conservative Party Tuesday night to succeed her as prime minister.
While adopting a gentler tone on the contentious issue of Britain's involvement in Europe, he shares her opposition to a single European currency and shares her belief in tight restraint on government spending.
FT 08 APR 93 / John Major's Year: Major's blue period - A year on from success at the polls, the prime minister's popularity has plunged.
The past 12 months have been hijacked by internal party differences over Europe, by the debacle surrounding UK withdrawal from the exchange rates mechanism of the European Monetary System, and by a continuing, deep recession which has disappointed and alienated many traditional Tory supporters in business.
Its Leader"] [Text] In local government elections across Britain yesterday, the Conservatives suffered their worst defeat ever, losing control of 17 regional councils and 444 seats.
Even before all of the results were known, some Tories openly announced their determination to challenge John Major's position and remove him from office as early as possible.

The extract generated by system 28 has 8 sentences of which only one of them contained relevant information. When comparing using sentence precision and recall, all three extracts only have one sentence match which is the first sentence. If we calculate the *F*-factors based on the model extract shown above, system 28 has a score of 0.143 and system 29 has a lower score of 0.118. After reading all three extracts, the extract generated by system 29 contains much more relevant information than that generated by system 28. The missing information in system 28 is ---*John Major and the Conservatives were losing the popularity in 1993, after John Major won the election one year ago*,-- which should be the most important content in the extract. In our DG-based approach, the scores assigned to system 28 and 29 are 0.063 and 0.100, respectively; which points out that systems 29 did a better job than system 28.

System rank	Sentence-based Ranking	Sentence-based *F*-factor	DG-based Ranking	DG-based *F*-factor
1 (worst)	22	0.000	28	0.092
2	16	0.062	22	0.101
3	31	0.081	16	0.111
4	25	0.081	20	0.115
5	29	0.090	25	0.115
6	20	0.125	21	0.122
7	28	0.138	31	0.124
8	24	0.171	24	0.125
9	19	0.184	19	0.129
10 (best)	21	0.188	29	0.132

Table 3: Machine-generated summaries vs. source documents. Ranking results for 200 words extracts

System rank	Sentence-based Ranking	Sentence-based *F*-factor	DG-based Ranking	DG-based *F*-factor
1 (worst)	22	0.000	22	0.137
2	16	0.128	28	0.141
3	25	0.147	16	0.160
4	31	0.150	25	0.163
5	29	0.155	20	0.164
6	20	0.172	31	0.165
7	28	0.197	21	0.167
8	24	0.223	29	0.168
9	19	0.224	19	0.168
10 (best)	21	0.258	24	0.169

Table 4: Machine-generated summaries vs. source documents. Ranking results for 400 words extracts

200-word		400-word	
System	*F*-factor	System	*F*-factor
28	0.092	22	0.137
22	0.101	28	0.141
16	0.111	16	0.160
20	0.115	25	0.163
25	0.115	20	0.164
21	0.122	31	0.165
Model	0.124	Model	0.165
31	0.124	21	0.167
24	0.125	29	0.168
19	0.129	19	0.168
29	0.132	24	0.169

Table 5: Average *F*-factors for the model summaries and machine-generated summaries.

Of the 59 submitted 200-word extracts by system 28, 39 extracts suffer the problem of having less informative sentences. The number of such sentences is 103, where the total number of sentences is 406 from all the extracts for system 28. On average, each extract contains 1.75 such sentences, where each extract has 6.88 sentences. For the 400-words extracts, we found 54 extracts among the 59 submitted summaries also have this problem. The total number of such sentences was 206, and the total number of sentences was 802 sentences. So, about 3.49 sentences do not contain much information, where the average length of each extract is 13.59 sentences. Thus, a large portion of each extract does not contribute to the do example, will not be considered a good summary, either on the criterion of summary coherence or summary informativeness, where coherence is how the summary reads and informativeness is how much information from the source is preserved in the summary (Mani, 2001).

From the results based on comparing extracts against original documents, we found that several systems perform very similarly, especially in the experiments with 400-word extracts (Table 4). The results show that except for systems 22 and 28 which perform significantly worse, all other systems are very similar, from the point of view of informativeness.

Finally, we generated DGs for the model extracts and then compared them against their original documents. The average *F*-factors are calculated, which are listed in Table 5 along with the scores for different automatic summarization systems. Intuitively, a system provides extracts that contain more information than other systems will get a higher score. As we can see from the data, at 200-words, the extracts generated by systems 21, 31, 24, 19, and 29 contain roughly the same amount of information as those created by humans, while the other five systems performed worse than human judges. At 400-words, when the compression ratio of the extracts is decreased, more systems perform well; only systems 22 and 28

generated summaries that contain much less information than the model summaries.

5 Discussion and Future Work

In DUC 2002 data collection, 9 human judges were involved in creating model extracts; however, there are only 2 model extracts generated for each document set. The sentence precisions and recalls obtained from comparing the machine generated extracts and human generated model extracts are distributed along with raw data (DUC-2002. http://www-nlpir.nist.gov/projects/duc), with the intent to use them in system performance comparison. Van Halteren (2002) argued that only two manually created extracts could not be used to form a sufficient basis for a good benchmark. To explore this issue, we obtained a ranking order for each human judge based on the extracts he/she generated. The results showed that the ranking orders obtained from 9 different judges are actually similar to each other, with the average Spearman correlation efficient to be 0.901. From this point of view, if the ranking orders obtained by sentence precision and recall based on the model extracts could not form a good basis for a benchmark, it is because of its binary nature (Jing et al., 1998), not the lack of sufficient model extracts in DUC 2002 data.

Van Halteren and Teufel (2003) proposed to evaluate summaries via *factoids*, a pseudo-semantic representation based on atomic information units. However, sufficient manually created model summaries are need; and *factoids* are also manually annotated. Donaway et al. (2000) suggested that it might be possible to use content-based measures for summarization evaluation without generating model summaries. Here, we presented our approach to evaluate the summaries base on document graphs, which is generated automatically. It is not very surprising that different measures rank summaries differently. A similar observation has been reported previously (Radev, et al, 2003). Our document graph approach on summarization evaluation is a new automatic way to evaluate machine-generated summaries, which measures the summaries from the point of view of informativeness. It has the potential to evaluate the quality of summaries, including extracts, abstracts, and multi-document summaries, without human involvement. To improve the performance of our system and better represent the content of the summaries and source documents, we are working in several areas: 1) Improve the results of natural language processing to capture information more accurately; 2) Incorporate a knowledge base, such as WordNet (Fellbaum, 1998), to address the synonymy problem; and, 3) Use more heuristics in our relation extraction and generation. We are also going to extend our experiments by comparing our approach to content-based measure approaches, such as cosine similarity based on term frequencies and LSI approaches, in both extracts and abstracts.

6 Acknowledgments

This work was supported in part by the Advanced Research and Development Activity (ARDA) U.S. Government. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U. S. Government.

References

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6): 391-407.

Robert L. Donaway, Kevin W. Drummey, and Laura A. Mather. 2000. A comparison of rankings produced by summarization evaluation measures. In *Proceedings of the Workshop on Automatic Summarization*, pages 69-78.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database.* MIT Press.

Jade Goldstein, Mark Kantrowitz, Vibhu Mittal, and Jaime Carbonell. 1999. *Summarizing text documents: Sentence selection and evaluation metrics*. In Proceedings the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 121-128, ACM, New York..

Hans van Haltern. 2002. Writing style recognition and sentence extraction. *DUC'02 Conference Proceedings.*

Hans Van Halteren and Simone Teufel, 2003. Examining the consensus between human summaries: Initial experiments with factoid analysis. In *HLT/NAACL-2003 Workshop on Automatic Summarization*.

Hongyan Jing, Kathleen McKeown, Regina Barzilay, and Michael Elhadad. 1998. Summarization evaluation methods: experiments and analysis.

In *American Association for Artificial Intelligence Spring Symposium Series*, pages 60-68.

Chen-Yew Lin. 2001. Summary evaluation environment. http://www.isi.edu/~cyl/SEE.

Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the Workshop on Automatic Summarization post conference workshop of ACL-02*, Philadelphia, PA, U.S.A., July 11-12 (DUC 2002).

Inderjeet Mani. 2001. Summarization evaluation: An overview. In *Proceedings of the NTCIR Workshop 2 Meeting on Evaluation of Chinese and Japanese Text Retrieval and Test Summarization*, National Institute of Informatics.

Inderjeet Mani and Mark T. Maybury. 1999. *Advances in Automatic Text Summarization*. The MIT Press.

Manuel Montes-y-Gómez, Alexander Gelbukh, and Aurelio López-López. 2000. Comparison of conceptual graphs. In *Proceedings of MICAI-2000, 1st Mexican International Conference on Artificial Intelligence*, Acapulco, Mexico.

Jerome L. Myers and Arnold D. Well. 1995. *Research Design and Statistical Analysis*, pages, 488-490, Lawrence Erlbaum Associates, New Jersey

Paul Over and Walter Liggett. 2002. Introduction to DUC-2002: An intrinsic evaluation of generic news text summarization systems. *Document Understanding Conferences website* (http://duc.nist.gov/)

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. *BLEU: A Method for Automatic Evaluation of Machine Translation*. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center.

Dragomir R. Radev, Simone Teufel, Horacio Saggion, Wai Lam, John Blitzer, Hong Qi, Arda Celebi, Danyu Liu, and Elliott Drabek. 2003. Evaluation challenges in large-scale document summarization. In *Proceedings of the 41st Annual Meetring of the Association for Computational Linguistics*, July 2003, pages 375-382.

Keith van Rijsbergen. 1979. *Information Retrieval*. Second Edition Butterworths, London.

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

Eugene Santos Jr., Hien Nguyen, and Scott M. Brown. 2001. Kavanah: An active user interface *Information Retrieval Agent Technology. Maebashi, Japan, October 2001*, pages 412-423.

Danny Sleator and Davy Temperley. 1993. Parsing English with a link grammar. In *Proceedings of the Third International Workshop on Parsing Technologies*, pages 277-292.

Karen Sparck-Jones and Julia R. Galliers. 1996. *Evaluating Natural Language Processing Systems: An Analysis and Review* (*Lecture Notes in Artificial Intelligence 1083*). Springer-Verlag