Automatic Checking of Alternative Texts on Web Pages.

Morten Goodwin Olsen[1], Mikael Snaprud[1,3] and Annika Nietzio[2]

[1] Tingtun AS,
PO Box 48, N−4791 Lillesand, Norway.
morten.g.olsen@tingtun.no, mikael.snaprud@tingtun.no
http://www.tingtun.no
[2] Forschungsinstitut Technologie und Behinderung (FTB)
der Evangelischen Stiftung Volmarstein, Grundschötteler Str. 40
58300 Wetter (Ruhr), Germany
egovmon@ftb-net.de
http://www.ftb-net.de
[3] University of Agder,
PO Box 422, N−4604 Kristiansand, Norway.

Abstract. For people who cannot see non-textual web content, such as images, maps or audio files, the alternative texts are crucial to understand and use the content. Alternate texts are often automatically generated by web publishing software or not properly provided by the author of the content. Such texts may impose web accessibility barriers. Automatic accessibility checkers in use today can only detect the presence of alternative texts, but not determine if the text is describing the corresponding content in any useful way. This paper presents a pattern recognition approach for automatic detection of alternative texts that may impose a barrier, reaching an accuracy of more then 90%.

1 Introduction

The Unified Web Evaluation Methodology (UWEM) [1,2] has been presented as a methodology for evaluating web sites according to the Web Content Accessibility Guidelines [3]. The UWEM includes both tests which can be applied manually by experts and tests which can be applied automatically by measurement tools and validators.

All automatic tests in the UWEM are deterministic, which has some drawbacks. As an example, one of the automatic UWEM tests checks whether an image (`<img>` element) has an alternative text. There are no automatic tests checking the validity of such alternative texts. This means, for a web site to conform to the automatic UWEM tests, any alternative text is sufficient. People and applications such as search engines, who are unable to see images, rely on the alternative text to convey the information of non-textual web content. If this information is not present, or when the text does not describe the image well, the information conveyed in the image is lost to these users.

To make sure web sites are accessible, appropriate textual alternatives are needed in many places, such as in frame titles, labels and alternative texts of graphical elements [3,4,5].[4] In many cases, these alternative texts have either been automatically added by the publishing software, or a misleading text has been supplied by the author of the content. Examples of such include alternative texts of images such as "Image 1", texts which resemble filenames such as "somepicture.jpg" or "insert alternative text here". Most automatic accessibility checkers, including validators that comply with the automatic UWEM tests, check only for the existence of alternative texts. The above mentioned texts, which are *undescriptive* and are thus not considered accessible, will not be detected by those tests. Our data shows that 80% of the alternative texts are not describing the corresponding content well.

This paper proposes an extension of UWEM with tests for automatic detection of alternative texts which, in its context, is in-accessible using pattern recognition algorithms.

To the best of our knowledge using pattern recognition to test for *undescriptive* use of alternative texts in web pages has not been done previously. However, similar related approaches has been conducted. For example, the Imergo Web Compliance Manager [6] provides results for suspicious alternative texts for images. The algorithm is not presented in the literature.

Furthermore, a technique for automatic judging of alternative text quality of images has been presented by Bigham [7].[5] This approach judges alternative texts in correspondence with the images, including classification using common words found in alternative texts and check if the same text is present in any other web page on Internet. The classifier uses Google and Yahoo, and has in their best experiment an accuracy of 86.3% using 351 images and corresponding alternative texts. However, in the presented results 7999 images are discarded because the algorithm fails to label the images. It is evident that discarding 7999 images (95.8%) is undesirable and has a severe impact on the over all accuracy. Taking the discarded images into account, the true accuracy of the presented algorithm is only 3.6%.

Detecting *undescriptive* texts in web pages has many similarities with detection of junk web pages and emails where heuristics and pattern classification has been successfully applied [8,9,10,11].

It is worth noticing that descriptiveness of a text could be seen in correspondence with the content. For example, if there is an image of a cat, an appropriate alternative textual description may be "cat" while "dog" would be wrong. In order to detect these situations image processing would most likely be needed in addition to text classification. Even though this could increase the over all accuracy, it is a much more challenging task which also includes significant increase

---

[4] All types of textual alternatives are in this paper referred to as alternative texts. An alternative text is *descriptive* if it describes the corresponding content well, and *undescriptive* if it does not.

[5] Note that this is limited only to alternative texts of images, while our approach includes several types of alternative texts.

in computational costs [12]. Image processing is not within the scope of this paper.

2   Approach

This paper presents a method for detecting *undescriptive* use of alternative texts using pattern recognition algorithms. The paper follows traditional classification approach [13]: Section 3 presents the data used for the classification. From this data features are extracted in section 4. The classification algorithms and results are presented in section 5; Naïve Bayes in section 5.1 and Nearest Neighbor in section 5.2. Finally, section 6 and 7 present the conclusions and further work.

3   Data

The home page of 414 web sites from Norwegian municipality were downloaded. From these, more then 11 000 alternatives texts were extracted (more than 1700 unique alternative texts) and manually classified as either:

- **Descriptive**: The alternative texts describes the corresponding content well and imposes no accessibility barrier.
- **Undescriptive**: The alternative texts does not the correspond to the content well and is a potential accessibility barrier.

All web pages have been deliberately chosen to be from only one language to avoid possible bias due to language issues such as; the length of words, which is language dependent [14], or words that are known to be *undescriptive* which will differ between languages [15]. Despite only using web pages from one language in this paper, the algorithms presented are not expected to be limited to only Norwegian. The algorithms can be applied to any language as long as appropriate training data is used.

Note that frequent problem of absence of *descriptive* texts [16] has deliberately been removed from these experiments. Testing for the presence of such *descriptive* texts are already present in UWEM as a fully automatable test [1,2] and is thus not addressed in this paper.

4   Feature Extraction

Several features of *undescriptive* texts in web pages have already been presented in the literature [7,15].

Slatin [7] found that file name extensions, such as .jpg, is a common attribute of *undescriptive* alternative texts of images. Additionally, the study indicates that a dictionary of known *undescriptive* words and phrases can be useful for such a classification.

Craven [15] presented additional features that characterizes *undescriptive* alternative texts. Most significantly, he found certain words/characters that are

common in *undescriptive* alternative texts such as "*", "1", "click", "arrow" and "home". Additionally, he found a correlation between the size of the image and length of the alternative text. His empirical data indicates that images of small sizes are more often used for decoration and should because of this have an empty alternative text.

In line with literature [7,5,15], the following features where extracted from the collected data:

- Number of words in the alternative text.
- Length of the alternative text (number of characters).
- File type abbreviations such as GIF, JPEG and PDF.
- None alphabet characters such as *, 1, ..
- Words that are known to cause accessibility barriers such as "read more", "click here", "title".[6]
- The size of the image presented in the alternative text (INTxINT).
- HTML such as .[7]

Generally speaking, features will work well as part of classifiers as long as they have a discriminatory effect on the data [13]. This means, based on the properties of the features alone, it should be possible to separate the data which belongs to both the *descriptive* and *undescriptive* classes.

The features have different properties and distributions. The features "number of words" and "length of the alternative text" are represented by positive integer values $(1, 2, 3, ...)$. The discriminatory effect of these features are presented as density graphs in figure 1. Figure 1 shows that common properties for the *undescriptive* texts are shorter alternative texts and fewer words. Most noticeably, figure 1 shows that having only one word in the alternative texts is a common property of the *undescriptive* class, while having two or more words is a common property for the *descriptive* class.

The remaining features are represented by a boolean value. As an example, a file name extension is either present or not present in the alternative texts. Figure 2 shows the discriminatory effect of features represented by boolean values.[8] As an example, close to 50% of the *undescriptive* alternative texts had words which often cause accessibility barriers, while only 0.5% of the *descriptive* alternative texts had the same behaviour. Similarly about 2% of the *undescriptive* texts had file name extensions, while only 0.05% *descriptive* texts had filename extentions.

5 Classification

Essential for the algorithms is the actual classification. In this paper we have implemented and tested two well known classification algorithms; Nearest Neighbor and Naïve Bayes [17,13].
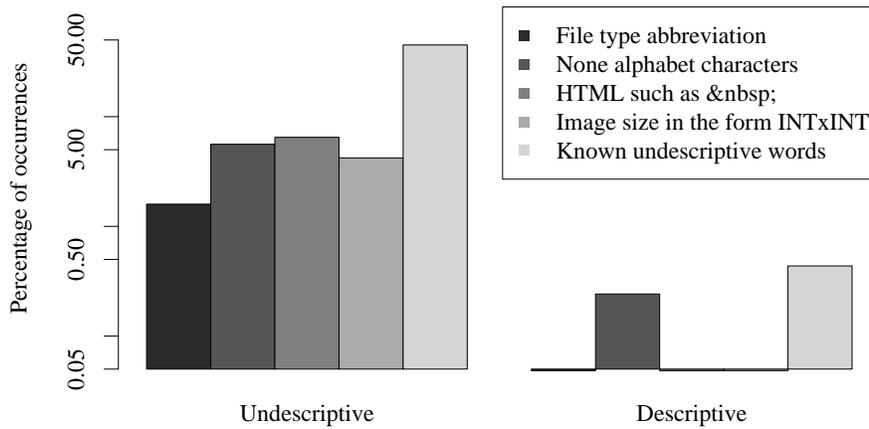
---

[6] Norwegian translations of these words were used.

[7] In this study, all types of HTML is included. It is worth noticing that not every entity is problematic.

[8] Note that the y-axis is logarithmic
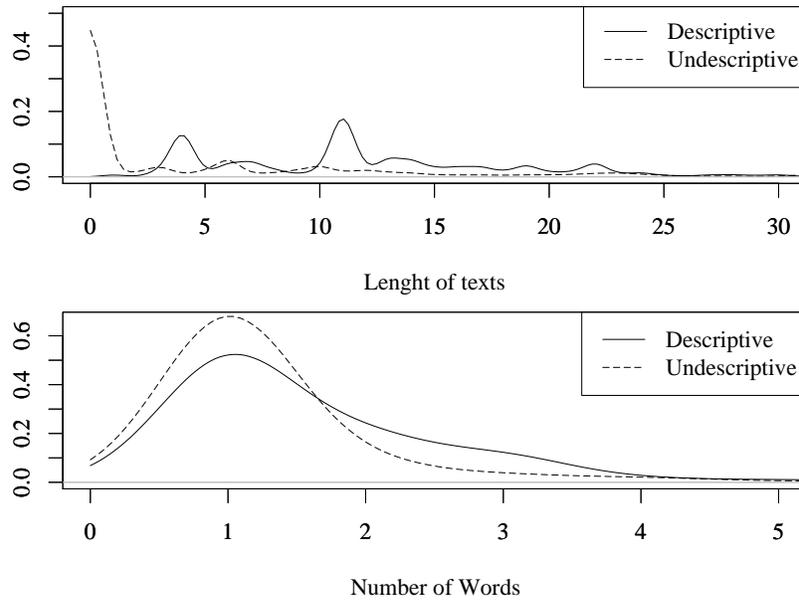
Fig. 1. Density Graphs for the features number of words and length of texts.

Fig. 2. Bar Charts with percentage of occurrence for features.

All algorithms have been tested with leave one out cross validation [18]; 1. Train with all data set except one instance. 2. classify the remaining instances. 3. Select next instance and go to 1. This ensures that the training sample is independent from test set.

5.1 Approach 1 - Nearest Neighbor

With the Nearest Neighbor algorithm, the data are added in multidimensional feature space where each feature represents a dimension, and every record is represented by a coordinate in the feature space. The euclidean distance is calculated between the item to be classified, and all items part of the training data. This identifies the nearest neighbors, and voting between the k nearest neighbors decides the outcome of the classification. In our experiments, k was chosen to be 1.

Figure 1 suggests that length of the alternative texts and number of words could be sufficient features for a working classifier. However, the empirical results does not support this as using these features alone gives the classifier an accuracy of only 66.5% and 69.0%.

By using all features described in section 4 the classifier achieves an accuracy of 90.0%, which is significantly higher than the state-of-the-art [7]. A confusion matrix with the classification results can be seen in table 1.

Table 1. Confusion Matrix with classification accuracy using Nearest Neighbor

	descriptive	undescriptive	over all accuracy
descriptive	93.9%	6.1%	
undescriptive	27.2%	72.8%	
over all accuracy			**90.0%**

5.2 Approach 2 - Naïve Bayes

How well the classification is working is dependant on how well the features are able to discriminate the classes. It could be that the chosen features described in section 4 are not the best way to identify *descriptive* and *undescriptive* alternative texts. The words themselves could have a significant discriminatory effect [19].

By relying on the words alone using a Naïve Bayes classifier, we get an accuracy of 91.9%. This is slightly more than Approach 1 and again significatly higher than the state-of-the-art [7]. A confusion with the classification results can be seen in table 2.

Table 2. Confusion Matrix with classification results using Naïve Bayes

	descriptive	undescriptive	over all accuracy
descriptive	92.6%	7.4%	
undescriptive	10.9%	89.1%	
over all accuracy			**91.9%**

6 Conclusion

This paper presents an approach for classifying alternative texts in Web pages as *descriptive* or *undescriptive*. *Undescriptive* texts are not describing the corresponding content well and may impose an accessibility barrier. The paper presents two approaches; classification based on well known properties of *undescriptive* texts presented in literature and classification using the texts alone. Both approaches have an accuracy of more then 90%, which is better than the state-of-the-art. Furthermore, in contrast to the state-of-the-art, this paper presents approaches that are independent from third party tools.

The findings in this paper gives a strong indication that *undescriptive* alternative texts in Web pages can be detected automatically with a high degree accuracy.

7 Further work

Further work includes looking at alternative texts in comparison with the actual images. This would include adding image processing to improve the over all accuracy.

We could expect that the content of the alternative texts are related to the text of the page itself. We would like to explore to what extent *descriptive* text could be topicwise related content of the web page and how this can potentially part of the features.

This paper only presents an approach to detect *undescriptive* use of alternative texts. In the future we will explore how results from the tests can be incorporated with the existing UWEM framework.

Acknowledgements

---

[9] http://www.egovmon.no

References

1. Web Accessibility Benchmarking Cluster: Unified Web Evaluation Methodology (UWEM 1.2). Retrieved November 4th, 2009, from http://www.wabcluster.org/uwem1_2/ (2007)
2. Nietzio, A., Ulltveit-Moe, N., Gjøsæter, T., Olsen, M.G., Snaprud, M.: Unified Web Evaluation Methodology Indicator Refinement. Retrieved November 4th, 2009, from http://www.eiao.net/resources (2007)
3. World Wide Web Consortium: Web Content Accessibility Guidelines 1.0. W3C Recommendation 5 May 1999. Retrieved November 4th, 2009, from http://www.w3.org/TR/WCAG10/ (1999)
4. World Wide Web Consortium: Web Content Accessibility Guidelines (WCAG) 2.0. Retrieved November 4th, 2009, from http://www.w3.org/TR/REC-WCAG20-20081211/ (2008)
5. Slatin, J.: The art of ALT: toward a more accessible Web. Computers and Composition **18**(1) (2001) 73–81
6. Web Compliance Center of the Fraunhofer Institute for Applied Information Technology (FIT): Web compliance manager demo. Retrieved November 4th, 2010, from http://www.imergo.com/home (2009)
7. Bigham, J.: Increasing web accessibility by automatically judging alternative text quality. In: Proceedings of the 12th international conference on Intelligent user interfaces, ACM (2007) 352
8. Hershkop, S.: Behavior-based email analysis with application to spam detection. PhD thesis, Citeseer (2006)
9. Fetterly, D., Manasse, M., Najork, M.: Spam, damn spam, and statistics: Using statistical analysis to locate spam web pages. In: Proceedings of the 7th International Workshop on the Web and Databases: colocated with ACM SIGMOD/PODS 2004, ACM New York, NY, USA (2004) 1–6
10. Yu, B., Xu, Z.: A comparative study for content-based dynamic spam classification using four machine learning algorithms. Knowledge-Based Systems **21**(4) (2008) 355–362
11. Kolesnikov, O., Lee, W., Lipton, R.: Filtering spam using search engines. Technical report, (Technical Report GITCC-04-15, Georgia Tech, College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, 2004-2005)
12. Chen, Z., Wu, O., Zhu, M., Hu, W.: A novel web page filtering system by combining texts and images. In: WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, USA, IEEE Computer Society (2006) 732–735
13. Duda, R., Hart, P., Stork, D.: Pattern classification. Citeseer (2001)
14. Solorio, T., Pérez-Coutino, M., et al.: A language independent method for question classification. In: Proceedings of the 20th international conference on Computational Linguistics, Association for Computational Linguistics (2004) 1374
15. Craven, T.: Some features of alt text associated with images in web pages. Information Research **11** (2006)
16. Gutierrez, C., Loucopoulos, C., Reinsch, R.: Disability-accessibility of airlines Web sites for US reservations online. Journal of Air Transport Management **11**(4) (2005) 239–247
17. Han, J., Kamber, M.: Data mining: concepts and techniques. Morgan Kaufmann (2006)

18. Wikipedia: Cross-validation (statistics) — wikipedia, the free encyclopedia (2010) [Online; accessed 28-January-2010].
19. Duchrow, T., Shtatland, T., Guettler, D., Pivovarov, M., Kramer, S., Weissleder, R.: Enhancing navigation in biomedical databases by community voting and database-driven text classification. BMC bioinformatics **10**(1) (2009) 317