Classic Children's Literature - Difficult to Read?

Dolf Trieschnigg[1] and Claudia Hauff[2]

[1] DB group
[2] HMI group,
University of Twente, Enschede, The Netherlands
{trieschn,hauffc}@ewi.utwente.nl

Abstract. Classic children's literature such as *Alice in Wonderland* is nowadays freely available thanks to initiatives such as Project Gutenberg. Due to diverging vocabularies and style, these texts are often not readily understandable to children in the present day. Our goal is to make such texts more accessible by aiding children in the reading process, in particular by automatically identifying the terms that result in low readability. As a first step, in this poster we report on a preliminary user study that investigates the extent of the vocabulary problem. We also propose and evaluate a basic approach to detect such difficult terminology.

1 Introduction

Many classic works of children's literature like *Alice in Wonderland* are nowadays freely available thanks to initiatives such as Project Gutenberg[1] (PG), a digital library which contains mostly public domain books. Many of these books were written more than one hundred years ago. Due to the diverging vocabularies, style and wording of the texts, they are often not readily understandable to children today. Our goal is to breach this vocabulary gap in order to make these classic works more accessible to children today.

Consider, for instance, the extract from *The Three Musketeers* in Figure 1: underlined are (multi-word) terms that we hypothesize to be unknown by most children today. Though the extract is short, five words are easily recognizable as unusual in today's English. The table in Figure 1 gives an indication of the vocabulary mismatch between works of different time periods. We downloaded between 5 and 10 books from PG's children's literature category for each 25 year time period starting with 1775-1800 and ending with 1900-1925. The unigram term distribution over all books of a time period was then calculated. Reported is the Jensen-Shannon divergence (JSD) [3] between those distributions. The larger the divergence, the more the term distributions of different time periods differ. As can be expected, with increasing difference in time periods, the divergence in the vocabulary increases. While this experiment only considers unigrams and not multi-word terms (or even the style of writing), it already shows that on the single term level alone large differences occur.

[1] http://www.gutenberg.org/

P. Clough et al. (Eds.): ECIR 2011, LNCS 6611, pp. 691–694, 2011.

To investigate the extent of the vocabulary gap in children's literature, we performed a user study where subjects were asked to tag the terms in paragraphs of children's books they were unfamiliar with or considered unusual. Those tagged terms were then used as a gold standard and a mechanism was developed to automatically detect them. We have two specific applications in mind: (i) a system that takes digital books from PG or other sources as input and generates an enhanced book in which difficult terminology is automatically linked to a definition, and, (ii) a system that aims to acquaint children with older vocabulary. While it would also be possible to create a system where children would interact with every word they do not know, it requires a lot of user input (such as clicks) and makes the reading process more tedious, in particular, if the number of unknown words is high.

	1800	1825	1850	1875	1900	1925
1800	0	0.25	0.26	0.33	0.34	0.40
1825		0	0.25	0.34	0.35	0.42
1850			0	0.25	0.26	0.30
1875				0	0.25	0.29
1900					0	0.26
1925						0

> "The citizens always took up arms readily against thieves, wolves or scoundrels, often against nobles or Huguenots, sometimes against the king, but never against cardinal or Spain. [...] the citizens, on hearing the clamor, and seeing neither the red-and-yellow standard nor the livery of the Duc de Richelieu, rushed toward the hostel of the Jolly Miller."

Fig. 1. The table shows the divergence in term distributions derived from books written in various time periods (1800 indicates the time period 1775-1800, and so on). The book extract on the right is from *The Three Musketeers* by Alexandre Dumas père (written in 1844). Underlined are terms we consider mostly unknown to children today.

To the best of our knowledge, no existing work deals specifically with the automatic detection of hard terminology in classic books. More remotely related work can be found in in the area of readability metrics for texts as a whole [1,4] and text simplification [2].

This poster is organized as follows: in Sec. 2 we present our user study. The detection approach is outlined in Sec. 3. Future work is discussed in Sec. 4.

2 User Study

For the user study, ten children's books were selected, ranging in time from *Robinson Crusoe* (1719) to *The Adventures of Paddy the Beaver* (1917). From each book, three paragraphs were randomly selected. The paragraph length ranges from 134 to 268 words. We recruited 30 subjects (11 female, average age: 31.8) from different research groups of several universities. All subjects are non-native English speakers; twenty-five participants judged their knowledge of English between 3-5 on a 1-5 Likert scale (where 5 indicates excellent). In this preliminary study, we make the simplifying (and debatable) assumption that terms that are tagged by good non-native English speakers are likely to be unknown to young native English speakers as well. The subjects were asked to tag the terms they were either unfamiliar with or deemed "difficult, rare or unusual".

We deliberately kept the description vague to get an understanding of what kind of terms the subjects consider difficult. The subjects were asked to tag three or more paragraphs; the average number of tagged paragraphs per subject was 4.1 and every paragraph was tagged by at least three subjects.

A (multi-word) term is added to the gold standard if at least half of all subjects that were presented the paragraph tagged it. Slight differences in tagging (e.g. one subject tags *stoop*, another tags *stoop down*) were manually cleaned up. This resulted in 39% of all the terms tagged by one or more subjects to be included in the gold standard. This number suggests that the agreement between the subjects was not always high. Some subjects tagged more than others, some tagged whole sentences. In the latter case, the sentence structure was deemed unusual. On average, 3.9 terms ($\sigma = 2.1$) were tagged in each paragraph. The minimum number (1 tag) was found in a paragraph from *The Adventures of Paddy the Beaver* (written in 1917), the maximum number (10 tags) was found in a paragraph from *The Legend of Sleepy Hollow* (written in 1820). The percentage of unique terms tagged in a paragraph varied between 0.9% and 8.3%, the average being 3.4%. The percentage of tagged terms varied considerably for the different paragraphs of each book. We did not observe a significant correlation between the year of writing and the percentage of unknown terms in a paragraph. Examples of tagged terms are *heathens*, *bonnets*, *garments*, *vicarage*, *horse balls*, *hillock* and *oratory.*

3 Detecting Difficult Terminology

We implemented a basic algorithm for the detection of difficult terminology backed up by definitions in Wiktionary[2]. The dictionary was compiled as follows. Wiktionary entries, limited to English nouns, verbs, adjectives and adjectives were extracted from a dump of Wiktionary pages. Stopwords and entries belonging to manually composed lists of inappropriate categories (such as *English basic words*, and *Sex*) were removed. Plural nouns and different verb forms were linked to their singular and simple present tense form, respectively. The filtered dictionary consists of 261,243 unique terms with 340,577 definitions. As an indicator of the difficulty of the terms, we determined the inverse collection frequency (ICF) of a term t (which is treated as a set of one or more words): $ICF(t) = \min_{w \in t} log \frac{n}{\max(cf(w),1)}$, where n is the total number of words in the collection and $cf(w)$ is the number of times the word w appears in the collection. Our intuition was that more difficult terminology has a higher ICF. We used a dump of Simple Wikipedia and a selection of 114 Gutenberg books (mainly novels, written between 1719-1920) as two sources to determine ICF.

Difficult terminology is detected in a two step process. (i) Finding candidates: a candidate set of difficult terms is detected by scanning the text for terms in the dictionary. In case overlapping terms are detected in the same string, the longest term is preferred. (ii) Filtering: candidate terms with an ICF below a predetermined threshold (based on training data) are discarded.

The collection obtained from the preliminary user study was used to evaluate our algorithm. 93% of the tagged terms also appeared in our dictionary.

[2] `http://www.wiktionary.org/`

In particular, some hyphenated terms were missed. We evaluated the detection performance in terms of precision (P), recall (R) and F-measure (F1) with and without ICF filtering based on 11-fold cross-validation. The fold (3 paragraphs) was used for training the optimal threshold, yielding the highest F1, the remaining paragraphs were used for testing. The reported metrics indicate the average over 11 folds. As we expected, no filtering resulted in low precision (P:0.10, R:0.89, F1:0.17). ICF filtering clearly improved precision at a smaller loss of precision, resulting in an optimal F-measure of 0.46 (based on Wikipedia: P:0.37, R:0.66, F1:0.24; based on Gutenberg books: P:0.44, R:0.53, F1:0.46). Given the relatively low inter annotator agreement (only 39% of the terms was tagged by multiple subjects), we think this is a reasonable first performance. The false positives of the Simple Wikipedia filtering revealed its limited coverage of typical story words (verbs such as *remarked*, and *nodded*). In contrast, the ICF from the Gutenberg books sometimes falsely filters nowadays uncommon terms. A combination of filtering from both sources might overcome these errors. Further improvements might come from additional features, such as number of syllables in the terms and using multiple ICF thresholds for different parts of speech.

Finally, we applied the detection mechanism with the Gutenberg corpus on each of the ten children's books selected for our user study. The percentage of *unique* terms detected as difficult varied between 8.8% (*The Adventures of Paddy the Beaver*, 1917) and 30% (*Tarzan of the Apes*, 1914). In the latter case, the five most often occurring difficult terms are *cruiser*, *gorilla*, *locket*, *bugs*, *primeval.*

4 Summary and Future Work

In this poster we reported the results of a user study that investigates the amount of unknown/unusual vocabulary found in classic children's books. The results show that the vocabulary gap is significant and needs to be addressed in a system that attempts to make classic children's books accessible to young readers. Our attempt at an automatic detection mechanism yielded reasonably good results.

In this preliminary study we relied on non-native English speakers in lieu of young native speakers, assuming that both types of users would tag similar words. This assumption needs to be tested in a further user study with children. Such a study would also give insight into how large the vocabulary gap is for different age groups.

Acknowledgements. This research was supported by the Netherlands Organization for Scientific Research, NWO, grant 612.066.513.

References

1. Collins-Thompson, K., Callan, J.: Predicting reading difficulty with statistical language models. JASIST 56(13), 1448–1462 (2005)
2. De Belder, J., Moens, M.F.: Text simplification for children. In: Towards Accessible Search Systems Workshop, pp. 19–26 (2010)
3. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory 37(1), 145–151 (1991)
4. Petersen, S.E., Ostendorf, M.: A machine learning approach to reading level assessment. Computer Speech and Language 23(1), 89–106 (2009)