Ranking German Texts by Comprehensibility for Foreign Document Retrieval

Michail Vlachos
IBM Research - Zurich, Switzerland

Theodoros Lappas
University of California, Riverside

ABSTRACT

Assume the result of a search is a set of documents in a language other than the native language of a user. How can one rank these documents based on their perceived comprehensibility (i.e., from easier to most difficult)? Our work addresses this question by providing metrics that estimate how difficult or common are the words that comprise a document. We take special consideration of language *cognates*, that is, words that are similar in two languages, which can significantly affect the understanding of a foreign-language text. Our evaluations on German documents when addressed to native English speakers, indicate that the comprehensibility estimator of a document, as provided by our technique, outperforms existing readability measures.

A video demonstration can be found here:
`http://www.youtube.com/watch?v=jHiZQ90OLg4`

INTRODUCTION

The web nowadays consists of large amounts of multilingual texts with overlapping content (e.g., news portals, reviews, blogs, RSS feeds, etc.). A typical web search may return documents in a language non-native to the user. If a subset of search results provides approximately the same coverage of the topic, how can the foreign documents be ranked based on their language difficulty?

To infer the difficulty of a foreign document, we estimate the average difficulty or commonness of its terms. In many studies it has been noted that the "log mean of word frequency...had the highest correlation with text difficulty" [4]. We leverage the knowledge distilled in web-search engines to estimate the frequency of foreign words. A novel aspect of our approach is that we also consider the native language of the reader. This is mainly due to the presence of *cognates*, i.e., words that are similar in both meaning and form in two languages. Such visual similarities between words can significantly ease the task of a reader. Essentially, this work provides a methodology for *sorting* foreign documents based on the difficulty, taking into consideration the native language of the user. Applications of this technology are:

1) **Language-aware personalization of the Web**. Using our approach one can rank and present 'similar' foreign articles to a non-native speaker, based on the perceived comprehension of the article, i.e., from most basic to most advanced usage of the foreign language. Also, the proposed method can be used for deciding *when* to translate a foreign document (i.e., only when deemed very difficult).

2) **Online Bookstores and Education**. Imagine the case of an English-speaking reader interested in German literature books. Which one should he/she read based on one's reading and communication skills? In addition, our methodology can be adapted for recommending the most suitable foreign reading material with respect to the student's native language [2].

Works similar in spirit to this work have been presented by Ott [5] and Uitdenbogerd [8]. These consider readability of foreign documents for educational purposes. They primarily examine the combination of existing readability metrics. The work of [6] considers the topic of supervised readability prediction, but highly depends on labeled data, while our method requires no such provisions. Finally, approaches such as the Lexile framework for document readability [4], do not consider the native language of the reader and ignore the effect of 'cognativity', an aspect to which we take special heed in our approach.

OVERVIEW

The comprehensibility of a foreign document d with respect to a language L can be estimated by two factors: the structural *readability* (rd) and the *familiarity* (fam) of the vocabulary. Formally: $C(d, L) = w_1 \times fam(d, L) + w_2 \times rd(d)$, where the weights $w_1, w_2 \geq 0$ are application dependent. Notice that familiarity of terms (and hence comprehensibility) is a function of the target language L. For example, a German document is expected to have higher comprehensibility value when read by Dutch rather than by Italian people, due to the smaller linguistic divergence of these two languages.

a) The readability of a document captures the *structural difficulty* of the text: how lengthy or perplexed is the structure. This can be estimated using various readability measures, such as the Flesch Reading Ease (FRE) measure [3]. FRE relies on various heuristic weights which attempt to capture the idea in the spirit of Zipf's Law: more frequent words are likely to have fewer syllables. Our evaluations focus on German documents, so we employ an instance of the

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

SIGIR 2011 Workshop on Enriching Information Retrieval (ENIR 2011), July 28, 2011, Beijing, China..

Flesch measure with its weights adapted to German documents [1].

b) The familiarity of a document vocabulary assesses how likely it is that the vocabulary used is known to the user. We define the measure as a function of two indicators: *popularity* and *cognativity*. Popularity captures the frequency of the document terms in texts written in the language under consideration; intuitively, rare terms are less likely to be familiar to the user. Cognativity is a language-dependent metric and measures the degree of affinity of a document's terms with respect to the user's own native language. Understandably, such terms are easier to understand.

Popularity: Words in a document have a (prior) global frequency-based measure of *popularity*, indicating how frequently they appear in documents of a given language. When browsing through a foreign document, a non-native speaker is more likely to recognize a very popular token, rather than one which is rarely used. We utilize the *collective knowledge* of the web to estimate these priors. Specifically, we use the page count from the Google search engine as an estimate of term popularity. This allows us also to accommodate for the popularity of newly introduced terms (e.g., iPad), which cannot generally be captured had one used static text corpora.

Definition 1 (Popularity). *The popularity of a term t is computed as the fraction:*

$$pop(t) = |\{t' : count(t') < count(t), t' \in \mathcal{V}\}|/|\mathcal{V}|, \qquad (1)$$

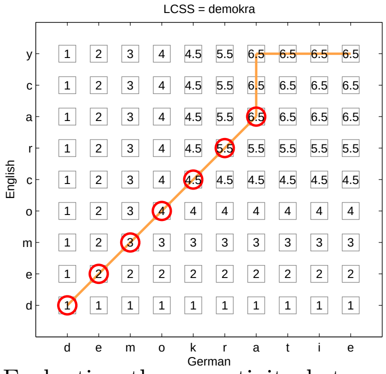
where $count(t)$ returns the number of appearances of a given token t in the entire document collection $\mathcal{D}$, and $\mathcal{V}$ is the vocabulary of all the distinct tokens in $\mathcal{D}$. The popularity of t is thus defined as the percentage of tokens in $\mathcal{V}$ that have fewer appearances in $\mathcal{D}$ than t. In addition to having a clear probabilistic interpretation, this formula is also robust to outliers (i.e., tokens with very low or very high frequencies).

Cognativity: Cognates are words in related languages that exhibit orthographic and semantic affinity. As an illustrative example, the German noun 'Haus' corresponds to the English word 'house'. Similarly the German adjective 'politisch' easily maps to 'political' in English. Identifying cognates in a text is important, because they affect bilingual language processing; presence of large number of cognates in a text can enhance its comprehensibility.

Using a bilingual dictionary, we employ a simple approach for spotting cognate words by exploiting the interlingual homography between a word and its translation. Our approach is based on a variation of the Longest Common Subsequence (LCSS). In particular, given a term t, let $\texttt{tr}(t, L)$ be its translation in the native language L of the user. Then, their similarity sim is defined as:

$$sim(t, \texttt{tr}(t, L)) = \frac{LCSS(t, \texttt{tr}(t, u))}{max(|t|, |\texttt{tr}(t, L)|)}$$

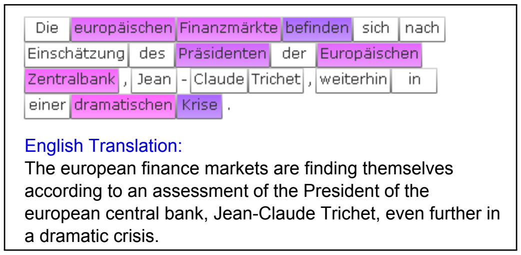
where $|\cdot|$ represents the length of a term. To better capture the letter transitions between the languages, fractional similarity values (e.g., 0.5) are assigned to dominant letter transfigurations. For example, the letter 'j' in German commonly maps to 'y' in English. As in 'jahr' → 'year'. Other dominant mappings are: 'k' → 'c' (e.g., architekt → architect) and 'z' → 'c' (e.g. sozial → social). We illustrate this in Figure 1, where we compute the distance between the German word 'demokratie' and its English translation 'democracy'. The final normalized similarity between the two words is $6.5/10 = 0.65$. Naturally, due to polysemy issues one needs to evaluate the similarity with all possible translations and retain the best score.



	d	e	m	o	k	r	a	t	i	e
y	1	2	3	4	4.5	5.5	6.5	6.5	6.5	6.5
c	1	2	3	4	4.5	5.5	6.5	6.5	6.5	6.5
a	1	2	3	4	4.5	5.5	6.5	6.5	6.5	6.5
r	1	2	3	4	4.5	5.5	5.5	5.5	5.5	5.5
c	1	2	3	4	4.5	4.5	4.5	4.5	4.5	4.5
o	1	2	3	4	4	4	4	4	4	4
m	1	2	3	3	3	3	3	3	3	3
e	1	2	2	2	2	2	2	2	2	2
d	1	1	1	1	1	1	1	1	1	1

Figure 1: Evaluating the cognativity between demokratie (German) and democracy (English)

Example: Below we demonstrate the cognate identification ability of our algorithm. Words with lighter (more red) background color have higher cognativity score compared to words in darker background.



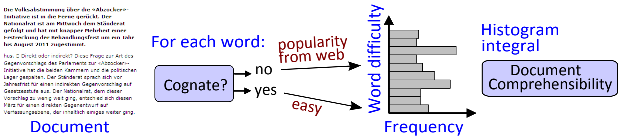
Die europäischen Finanzmärkte befinden sich nach Einschätzung des Präsidenten der Europäischen Zentralbank , Jean - Claude Trichet , weiterhin in einer dramatischen Krise .

English Translation:
The european finance markets are finding themselves according to an assessment of the President of the european central bank, Jean-Claude Trichet, even further in a dramatic crisis.

Combining Popularity and Cognativity: The identification of cognates is used in order to properly assess word familiarity, irrespective of its web popularity. In other words, cognativity is the dominant factor: if a term is the same (or almost the same) in the user's native language, then it is expected to be familiar even if the term is rarely used. We consider a word as a cognate if the cognativity value is greater than a cutoff threshold value ξ. (For our experiments we set $\xi = 0.45$, a value derived using a cross-validation on the results of a relevant user-study.) Cognates are assigned the maximum possible familiarity, equal to 1, while non-cognates are assigned a value equal to their popularity.

$$fam(t, L) = \begin{cases} pop(t), & \xi < cogn(t, L) \\ 1, & \xi \geq cogn(t, L) \end{cases} \qquad (2)$$

Equation 2 provides the familiarity of a single term. The familiarity of a document is computed by creating the histogram of the familiarity scores of all the terms in the document and then evaluating the *integral* of the histogram. This process is depicted graphically in Figure 2.



Die Volksabstimmung über die «Abzocker»-Initiative ist in die Ferne gerückt. Der Nationalrat ist am Mittwoch dem Ständerat gefolgt und hat mit knapper Mehrheit einer Erstreckung der Behandlungsfrist um ein Jahr bis August 2011 zugestimmt.

hus. ▫ Direkt oder indirekt? Diese Frage zur Art des Gegenvorschlags des Parlaments zur «Abzocker»-Initiative hat die beiden Kammern und die politischen Lager gespalten. Der Ständerat sprach sich vor Jahresfrist für einen indirekten Gegenvorschlag auf Gesetzesstufe aus. Der Nationalrat, dem dieser Vorschlag zu wenig weit ging, entschied sich diesen März für einen direkten Gegenentwurf auf Verfassungsebene, der inhaltlich einiges weiter ging.

Figure 2: Computing the overall document comprehensibility from the individual word scores

APPENDIX

Word Decompounding: We briefly touch upon the topic of work decompounding, which is important in our setting. Several languages such as German, Dutch or Swedish, are known as *compounding languages* because they allow the creation of new complex words by merging together simpler ones. It is therefore important to identify compound words and evaluate the individual familiarity of their components. This is because even though the compound word might be rare, if comprised by common words, then its meaning also becomes apparent. Schiller identified more than 40% of the words in a large German newspaper corpus as compounds [7]. Examples of German compounds are: Aschewolke (=ash clouds), sozialdemokratie (=social democracy).

Our methodology effectively identifies 2- and 3-compounds, by discovering the split point in a word that maximizes the popularity of the two subcomponents. The final familiarity of the word is the maximum between the decompounded familiarity and the original one (when treating the word as a whole). Illustratively, in the texts of the previous section, the words Finanzmärkte and Zentralbank had higher familiarity when decompounded rather than when considered as non-compound words.

Experiments: We illustrate the ability of our approach to capture the inherent comprehensibility of foreign textual content. We focus on German texts. We use data provided by the educational website `CourseInfo.com`, which hosts essays on a variety of topics, including foreign languages. The site provide reading material of variable difficulty levels for students native in English. Essays are grouped into 3 levels of increasing difficulty: GCSE (300 essays for high school students), A-level (150 essays for pre-college preparation) and University-level (50 essays for Bachelor-level students). In our experiment, we use all the available essays from the "German Essays" category.

For the first part of our evaluation, we use our approach to measure the comprehensibility of each essay, using an equal weight for readability and familiarity. As mentioned above, each essay belongs to one of three difficulty levels: `A-Level`, `GCSE` or `University`. Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be the sets of essays corresponding to two of the three levels and assume that $\mathcal{D}_1$ corresponds to a level easier than D_2 (e.g. $\mathcal{D}_1$ has the essays from `GCSE` and $\mathcal{D}_2$ from `University`). Then, the observed error percentage for this pair is:

$$error(\mathcal{D}_1, \mathcal{D}_2) = \frac{|\{(d_1, d_2) : d_1 \in \mathcal{D}_2, d_2 \in \mathcal{D}_2, C(d_1) < C(d_2)\}|}{|\mathcal{D}_1| \times |\mathcal{D}_2|} \tag{3}$$

The error is defined as the fraction of possible essay-pairs (d_1, d_2), where $d_1 \in \mathcal{D}_1$ and $d_2 \in \mathcal{D}_2$ and d_1 has received a lower comprehensibility score by our approach than d_2. This is undesirable, since Eq. 3 assumes that $\mathcal{D}_1$ corresponds to an easier level than $\mathcal{D}_2$. The computed error values for all possible level-combinations are shown in Table .

Table 1: Observed error for CourseInfo Data

Confusion Matrix		
Levels	A-level	University
GCSE	13.7%	3.1%
A-level		27.5%

Observe that for `GCSE` and `University` (the two levels that differ the most in terms of difficulty) the observed error was minimal (3.1%). A small error was also observed for the `GCSE` and `A-Level` pair, indicating that our approach can consistently distinguish `GCSE` essays. The highest error was observed for the `A-Level`/`University` pair. An inspection of some of the erroneous pairs revealed that deducing the true level of difficulty was an ambiguous task, even for a human annotator. Still, as shown in the table, such pairs were less than a third of the total. In short, our approach performed consistently well, managing to detect the, often subtle, gap in comprehensibility among the three levels.

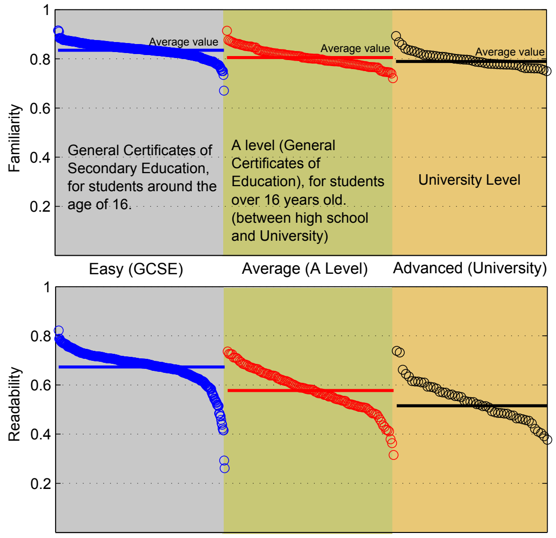


Figure 3: Comparing Familiarity and Readability. Familiarity is a more robust estimator of the document's difficulty.

Readability vs Familiarity: Finally, we demonstrate that the proposed familiarity measure is a more robust estimator of a document's comprehensibility. Figure 3 plots the readability and familiarity of the 3 classes of documents used in the previous experiments, in descending order. Even though both measures can provide accurate class distinction with respect to the average document value per class, familiarity is clearly a more *robust* estimator, because it introduces very little in-class variance. In general, as described in the previous sections, according to the application at hand, it is instructive to merge the two measures. Both measures offer a different view of a document's difficulty. However, when addressing foreign documents more weight should be given on the vocabulary aspect of a document, which is crystallized in the proposed familiarity measure.

REFERENCES

[1] T. Amstad. Wie verständlich sind unsere zeitungen? In *Universität Zürich: Dissertation*, 1978.

[2] T. Bell. Extensive reading: speed and comprehension. In *The Reading Matrix, 1(1)*, 2001.

[3] W. DuBay. The principles of readability. In *Impact Information*, 2004.

[4] C. Lennon and H. Burdick. The lexile framework as an approach for reading measurement and success. 2004.

[5] N. Ott. Information retrieval for language learning: An exploration of text difficulty measures. In *Master's Thesis in Computational Linguistics, Universität Tübingen*, 2009.

[6] S. E. Petersen. Natural language processing tools for reading level assessment and text simplification for bilingual education. In *Doctoral Thesis, University of Washington.*, 2007.

[7] A. Schiller. German compound analysis with wfsc. In *Finite-State Methods and Natural Language Processing*, 2006.

[8] A. L. Uitdenbogerd. Web readability and computer-assisted language learning. In *Proc. of Australasian Language Technology Workshop*, pages 99–106, 2006.