Customizing Search Results for Non-Native Speakers

Theodoros Lappas
Boston University
tlappas@cs.bu.edu

Michail Vlachos
IBM Research - Zurich

ABSTRACT

Blog posts, news articles and other webpages are present on the web in multiple languages. Presently, search engines primarily focus on *relevance* search with respect to the given text query. However, when considering documents with overlapping content, many of them written in a foreign language other than the user's own native tongue, it is beneficial to promote 'easier' to read documents. Here, we show how to rank a collection of foreign documents based on both: a) relevance to the query, and b) the comprehension difficulty of the document. We design effective ranking operators that evaluate the difficulty of a foreign document with respect to the user's native language. We show that existing search engines can easily augment their scoring function by incorporating the proposed comprehensibility metrics. Finally, we provide extensive experimental evidence that the comprehensibility-aware ranking model significantly improves the standard relevance-based ranking paradigm.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search Process*; I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text analysis*

Keywords

multilingual document search, document comprehensibility

1. INTRODUCTION

Large numbers of texts discussing the same topic can nowadays be retrieved from Web sources around the world (e.g. news portals, reviews, blogs, RSS feeds, etc.). As a result, a typical web search may return similar documents in multiple languages. The question that we are addressing in this work is how to build an engine that delivers not only the most relevant documents, but also the ones that best match the user's comprehension level of a foreign language. Foreign documents that are easier to read and understand should be ranked higher than more advanced texts with the same coverage of the topic. Given a collection of foreign documents (e.g. books, articles, news articles), we provide a structured methodology to effectively and accurately rank them based on their estimated *comprehensibility*. We then use this mechanism to build a search engine that considers both relevance and comprehensibility when evaluating candidate documents. To the best of our knowledge, this is the first approach that examines the problem of document ranking through the prism of foreign language difficulty, using a completely unsupervised approach.

The problem is challenging because it lies at the confluence of fields as diverse as linguistics, information retrieval and machine learning. Our approach combines both structural and linguistic features, exploring the different aspects of document comprehensibility. An additional dimension that we consider when estimating the reading difficulty of a foreign document, is the native language of the reader. For example, for a native Portuguese speaker, it can be significantly easier to comprehend Spanish documents rather than documents written in Greek or German. 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. We incorporate the identification of such word instances in our methodology.

We envision numerous applications where our methodology can be of use:

1) **Customization of search results**. Given the user's personal linguistic skills, our methodology enables the multilingual personalization of a search session, by evaluating and ranking foreign documents based on their comprehensibility.

2) **Language learning.** Studies have suggested that learning a foreign language is more effective when studying texts that match one's comprehension level [1]. Our work can be used to recommend the most suitable reading material to foreign language students.

3) **Machine Translation.** By estimating the comprehensibility of a given document, we can determine whether it falls within the language skills of the reader, or whether a translation should be attempted.

2. OVERVIEW

Our methodology estimates the *comprehensibility* of foreign documents. It depends on two primary factors: *readability*, which assesses the structural features of a given document, and *familiarity*, which focuses on the vocabulary.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.

Each of these two components captures a different aspect of comprehensibility. An illustration of our mechanism for evaluating comprehensibility is shown in Figure 1. The com-

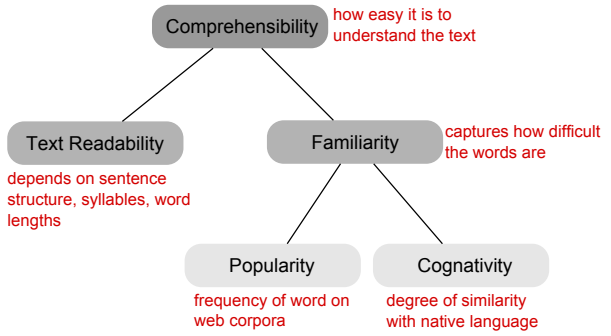

Figure 1: Document comprehensibility in our model

prehensibility of a document d with respect to a language L is defined as a linear combination of readability and familiarity. Formally:

$$C(d, L) = w_1 \times fam(d, L) + w_2 \times rd(d), \quad (1)$$

where $fam(d, L)$ denotes the familiarity of document d to user who is native (or proficient) in language L, and $rd(d)$ denotes the readability of d. Notice that familiarity (and hence comprehensibility) is defined as a function of the target language L. For example, a German document is expected to have higher comprehensibility value when read by Dutch people rather than by Italian people due to the higher linguistic similarity between German and Dutch. Finally, the two non-negative weights w_1 and w_2 are used to tune the impact of each factor.

3. FAMILIARITY

The familiarity of a document assesses how likely it is that its vocabulary is known to the user. We define the measure as a function of two indicators: *popularity* and *cognativity*. Popularity attempts to captures the general prevalence (i.e frequency) of terms in the language. Intuitively, rare terms are less likely to be familiar to the user. Cognativity is a language-dependent measure. Its use is to capture the degree to which a document's terms are similar in the user's own native language; normally, such terms would be easier to understand. Next, we discuss further these two factors.

3.1 Popularity

Intuitively, when reading a foreign document, a non-native speaker is more likely to recognize a very popular token than one which is rarely used. In a broader context, a document consisting of commonly used tokens is much easier to comprehend than another that uses more esoteric and unfamiliar vocabulary. In order to capture this "prior frequency" of a given a token t, we utilize the *collective knowledge* of the web. Today, most search engines provide the number of pages that the query appears in. We use this information as an estimate of term popularity[1]. An added advantage of using search engines instead of pre-existing text corpora, is the fact that online texts capture newly used terms, which is important since languages constitute an evolving organism.

[1]Specifically, we use the page count from Google.

Finally, search engines provide the functionality of focusing on a particular *language* for the documents to be retrieved.

Formally, *popularity* is defined as:

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

$$pop(t) = |\{t' : count(t') < count(t), t' \in \mathcal{V}\}|/|\mathcal{V}|, \quad (2)$$

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 robust to outliers (i.e., tokens with very low or very high frequencies) and serves as an intuitive and parameter-free way to smooth the obtained counts. Alternative smoothing techniques have been proposed in the literature [2].

3.2 Cognativity

Consider the following sentence: "Ein Experte kam die Maschine zu reparieren". A person proficient in English can easily deduce that this sentence translates to "An expert came to repair the machine", even if one is only a novice in German. The inherent familiarity of this sentence is due to the existence of *cognates*. Cognates are words in different languages that exhibit both orthographic and semantic affinity. In our work, we spot cognate words by exploiting interlingual homography. Our approach is based on the well-known problem of finding the Longest Common Subsequence (LCS) of two strings. In particular, given a term t, let $\texttt{tr}(t, L)$ be its translation in the native language L of the user. Then, we define their similarity sim as follows:

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

where $|\cdot|$ represents the length of a given string. Clearly, the measure assumes values in $[0, 1]$, evaluating the visual similarity between the term and its translation. Naturally, due to polysemy issues, we need to evaluate the term's similarity with all possible translations in the target language, and retain the best score. Let $\mathcal{T}(t, L)$ contain all translations of t in language L. Then, we define the cognativity of the term t with respect to L as:

$$cogn(t, L) = \max_{\texttt{tr}(t) \in \mathcal{T}(t, L)} sim(t, \texttt{tr}(t, L))$$

We consider a word as a cognate if its cognativity value is greater than a cutoff threshold value ξ. In our experiments, we set $\xi = 0.45$ which yielded the best results across languages. Terms identified as cognates are assigned the maximum possible familiarity (i.e. 1). The familiarity of non-cognates is equal to their popularity. Formally, we define the familiarity of a term t with respect to a language L as follows:

$$fam(t, L) = \begin{cases} pop(t), & \xi < cogn(t, L) \\ 1, & \xi \geq cogn(t, L) \end{cases} \quad (3)$$

Equation 3 gives us the familiarity of a single term. We define the aggregate familiarity of an entire document by

$$fam(d) \quad := \quad \sum_{t \in d} \frac{count(t, d)}{|d|} fam(t, L), \quad (4)$$

where $count(t, d)$ is the total number of appearances of term t in document d, and $|d| = \sum_{t \in d} count(t, d)$ is the total number of terms in d.

3.3 Word Decompounding

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. Schiller identified more than 40% of the words in a large German newspaper corpus as compounds [16]. As an example, the German compound word 'Medizindoktor' (=medical doctor) cannot be found in a dictionary and potentially also has few occurrences in texts or the web; however, its meaning is easily discernible given its building blocks. The splitting of a compound word in its basic parts is called *decompounding*. Our methodology is equipped with an effective algorithm for identifying 2- and 3-compounds. For ease of exposition, we provide next a method for the detection of 2-compounds. This method can be directly extended to identify 3-compounds.

Given a term t of length n, let $sub(\alpha, i, j)$ return the substring of α that begins at position i (inclusive) and ends at position j (inclusive). If $i > j$, the function returns $\emptyset$. We define the *decompounded* familiarity of t with respect to a language L as follows:

$$fam_{DC}(t, L) = \max_i \frac{1}{2}\{fam(sub(t, 1, i), u) + fam(sub(t, i+1, n), u)\}$$

where $1 \leq i \leq n$ and $fam(\emptyset) = 0$. Therefore, the above formula discovers the split point that maximizes the popularity of the two sub-components.

4. READABILITY

The notion of document readability has been a well-studied topic, particularly for English documents [18]. Many readability formulas have been proposed, all attempting to assign a single numerical readability score to each document. The most popular example is the Flesch Reading Ease (FRE) measure [10], which consists of a linear function of the mean number of syllables per word and the mean number of words per sentence in the document. The measure has been adapted to several languages, including English, French, Spanish, Italian and German[2]. For example, the formalization of the measure for German documents is:

$$FRE(d) = 180 - \frac{words(d)}{sents(d)} - 58.5 \times \frac{syllables(d)}{words(d)}, \quad (5)$$

where $words(d), sents(d)$ and $syllables(d)$ denote the number of words, sentences and syllables in d, respectively. The weights on the above formula have been derived by means of regression on training data. The Flesch Reading Ease yields numbers from 0 to 100, expressing the range from 'very difficult' to 'very easy', and is meant to be used for measuring the readability of texts addressed to adult language users. We choose FRE as a measure of readability, because of its popularity and widespread use as a readability yardstick in many organizations (e.g., U.S. Department of Defense).

Finally, we define the readability $rd(d)$ of a document d as the normalized version of $FRE(d)$, taken by dividing the score with the maximum $FRE(\cdot)$ observed over our entire collection $\mathcal{D}$. Formally:

$$rd(d) := \frac{FRE(d)}{\max_{d' \in \mathcal{D}} FRE(d')}, \quad (6)$$

5. SKYLINE RANKING

Next, we discuss how comprehensibility can be combined with relevance, toward a complete search engine for foreign-document retrieval. In our work, we define the relevance $rel(d)$ of a given document d via a combination of the Boolean Model and the Vector Space Model, as implemented in the popular Lucene search engine[3]. We further normalize the relevance values by dividing the score of each document with the maximum value observed over the entire corpus.

A document d can be represented by a two-dimensional vector $(C(d, L), rel(d)) \in \mathbb{R}^2_+$, where $C(d, L), rel(d)$ denote the document's comprehensibility (given the user's native language L) and relevance, respectively. We say that document d_1, dominates a document d_2 if d_1 is both more comprehensible and more relevant to the given query. Documents that are not dominated by any other document compose the *skyline* $\mathcal{S} \subset \mathcal{D}$ of the entire corpus $\mathcal{D}$.

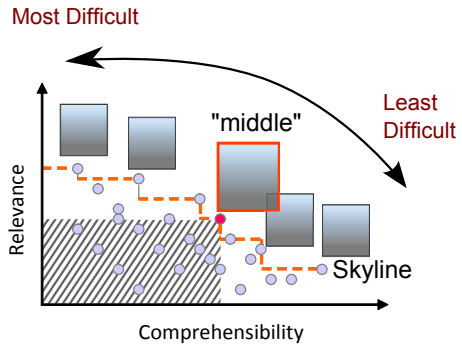

Figure 2: Navigating on the skyline of top-rated results

The skyline serves as an intuitive way to browse the promising documents of the search results. A good starting point is the skyline-document d^* that maximizes $(rel(d^*)+C(d^*, L))/2$. This point is annotated as 'middle' document in Figure 2. If the starting point is not comprehensible enough, the user is presented the next document to the right of the skyline. Note that the next point will be less relevant, otherwise it would dominate the point before it. Similarly, if the document not relevant enough, the next document to the left of the skyline is considered. Since the entire navigation process focuses on the skyline points, it is guaranteed to lead to document that is both comprehensible and relevant enough, if such a document exists. Further, this interface can assist in the evaluation of the relative comprehensibility and relevance of *any* document on the search result, based on its distance from the skyline points. We refer to this approach as `LingoRank`.

6. EXPERIMENTS

In this section we illustrate the ability of our approach to capture the inherent comprehensibility of foreign textual content. We start with a user study and then proceed to experiments on larger corpora.

[2] `http://www.ideosity.com/ideosphere/seo-information/readability-tests`

[3] `http://lucene.apache.org/java/3_0_2/api/all/org/apache/lucene/search/Similarity.html`

6.1 User Study

Initially, we want to estimate how well the proposed comprehensibility measure approximates the ranking provided by human annotators. We have assembled documents that address the same general topic but examine different aspects of it and possibly addressing different audiences. The topic we have focused on is the financial crisis in Greece (2010-2012). In order to include texts of variable comprehensibility, we have selected texts from sources with consistent language levels: 3 segments from financial websites (sophisticated and formal language with technical terms), 3 segments from mainstream news portals (edited, well-structured content with an average level of sophistication), and 3 segments from relevant comments posted in public forums (simpler, informal language). Nine German texts were given to eight human annotators who are native (or proficient) in English, but possess only a basic command of the German language. The same process was repeated for nine English texts, which were given to eight annotators native in German, with basic command of the English language.

The annotators were asked to rank the texts from easiest to most difficult. We also computed the scores for each text, using the developed comprehensibility formula. For this study, we used equal weights for familiarity and readability. The results are shown in Figure 6.1. The first column of each table shows the rank of each text based on the scores assigned by our method, the second and third columns hold the average rating and the standard deviation assigned from the annotators, respectively.

German Texts		
ours	user Avg	Std
1	1	0
2	2.8	0.9
3	3	0.82
4	3.4	1.3
5	5.9	1.25
6	6.7	1.28
7	6.9	0.99
8	6.3	1.28
9	9	0

English Texts		
ours	user Avg	Std
1	1.5	0.53
2	2.38	1.41
3	3.25	1.28
4	4.63	1.41
5	4.88	1.55
6	7.38	1.85
7	6.88	1.55
8	6.13	2.47
9	8	0.76

Figure 3: User Study on German and English texts: texts are ranked by our technique, as well as by human annotators.

The results of the study are very encouraging. For both German and English texts, the rank given by our method is consistently close to the average human rating. Our methodology was successful in ranking the texts by comprehensibility, illustrating its potential usefulness in the context of foreign document retrieval. For German documents, the observed standard deviation on the ratings was low, indicating a strong consensus among the annotators. The respective values for English were more elevated, suggesting that the same task on the English texts was more challenging. Nonetheless, our comprehensibility formula was still able to capture the average consensus rating of the annotators.

6.2 Large-Scale Evaluation on Real Data

In this experiment, we use data from the educational website `CourseInfo.com`, which hosts essays on a variety of topics, including foreign languages. On the website, 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). We use all available essays from the "German Essays" category.

First, we measure the comprehensibility of each essay. We tune the weights of familiarity and readability by minimizing the Minimum Squared Error (MSE) over the annotations of our user study. Specifically, the weights for familiarity and readability were set to 0.65 and 0.35, respectively. As mentioned above, each essay belongs to one of three difficulty levels: `A-Level`, `GCSE` or `University`. Given two different levels, we define the error to be the fraction of essay pairs that contain an essay from each level, such that the essay from the easier level received a lower comprehensibility score than the one from the higher level. The values for all possible level combinations are the following:

- `A-level` Vs. `GCSE` $\rightarrow$ 13.7%
- `A-level` Vs. `University` $\rightarrow$ 27.5%
- `GCSE` Vs. `University` $\rightarrow$ 3.1%

Observe that for `GCSE` and `University` (the two levels that differ the most in terms of difficulty) the observed error was very low (3.1%). A small error was also observed for the `GCSE` and `A-Level` pair, indicating that our approach can consistently distinguish `GCSE` essays. An inspection of the erroneous pairs for the `A-Level/University` pair 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 made up for less than a third of the total corpus.

6.3 Document Search

Here, we compare our methodology to standard relevance-based techniques. Specifically, we show that our approach allows the user to consider significantly fewer documents, before locating one that is both comprehensible and relevant to the given query. For this experiment, we collected a total of 1,002,394 articles from Google News, written in four different languages: German, Italian, Spanish, French. The corpus spans a time period of four months, between August 2011 and November 2011.

The experiment is performed independently for each language. First, we compose a list of 10 (foreign language) queries for each language, pertaining to major events that happened in the time-frame of the news articles. For this, we consult the list of important events of 2011, as reported in Wikipedia [4]. For each query, we retrieve the top-100 relevant documents, using Lucene's search functionality. For each document d in the top-100, we set d as the *target*. The target's comprehensibility and relevance values determine the lower bounds of the search: any document satisfying both bounds is considered a match. An approach is then evaluated based on how many documents it needs to present to the user until a match is found. We report the average number of examined documents for each approach (out of the 100). Since the process is repeated independently for each of the 10 queries, for each of the four languages, and for each of the documents in the top-100, we simulate a total of $4 \times 10 \times 100 = 4000$ search sessions. We use Lucene's relevance-based engine as a baseline. Starting from the most

[4] http://en.wikipedia.org/wiki/2011

relevant document, we traverse downward until a comprehensible document is reached. We then report the number of documents that had to be examines. We refer to this approach as `RelSort`, and to our own approach as `LingoRank`.

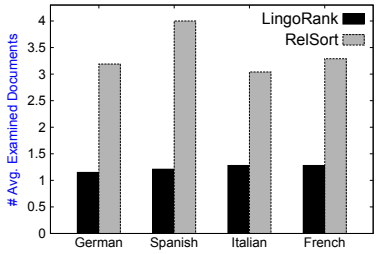

Figure 4: Number of documents that are considered by LingoRank and RelSort until a match is found.

The results of both approaches are shown in Figure 4. The y-axis shows the number of documents that had to considered until a match was found. We can see in the Figure that our approach outperformed `RelSort`, consistently considering fewer documents across all four languages. In fact, the average value observed for LingoRank was around one (1), suggesting that the starting point chosen by our approach (i.e., the document with the best comprehensibility/relevance mixture) was often the only one that needed to be considered.

7. RELATED WORK

Our work is related to the evaluation of textual *readability* [6]. However, our problem is far more rich and challenging, since we want to assess the comprehensibility of a *foreign* document. This depends not only on structural features, but also on linguistic features. Work on *text readability* can be broadly categorized into supervised and unsupervised. Unsupervised approaches rely on two aspects of text: the familiarity of the reader with its semantic units (words or phrases) and the complexity of its syntax. In order to define a metric for the former, linguistic resources ranging from manually compiled lists of words [3] to language models [5] have been employed. For syntactic complexity, the average sentence length is widely used, since it has been found to be strongly correlated with comprehensibility [7, 19, 9]. Supervised approaches exploit the availability of training data in order to derive statistical language models of readability for a particular language [14, 4, 15, 13]. Finally, approaches such as the Lexile framework for document readability [11], do not consider the native language of the reader and ignore the effect of cognativity.

Further, other relevant papers include studies on cognativity [8, 17], a concept that is a part of our own methodology. Finally, the impact of readability in the context of education and language learning has been explored by Ott [12].

8. CONCLUSION AND FUTURE WORK

In this work we described a search engine for foreign-document retrieval. The novelty of our engine lies with the consideration of a document's comprehensibility, in addition to its relevance to the given query. Our experimental evaluation verified the efficacy of our approach, and demonstrated its advantage when compared with standard techniques that focus exclusively on relevance.

9. REFERENCES

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

[2] T. Brants, A. Popat, P. Xu, F. Och, and J. Dean. Large language models in machine translation. In *Proc. of EMNLP-CoNLL*, 2007.

[3] J. Chall and E. Dale. *Readability revisited: The New Dale-Chall Readability Formula*. Brookline Books, 1995.

[4] K. Collins-Thompson and J. Callan. Predicting reading difficulty with statistical language models. In *JASIST '05*.

[5] K. Collins-Thompson and J. Callan. A language modeling approach to predicting reading difficulty. In *Proc. of HLT/NAACL*, 2004.

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

[7] R. Flesch. A new readability yardstick. In *J Appl Psychol 32*, pages 221–224, 1948.

[8] B. M. Friel and S. M. Kennison. Identifying German-English cognates, false cognates, and non-cognates: methodological issues and descriptive norms. In *Bilingualism: Language and Cognition '01)*.

[9] M. Heilman, K. Collins-Thompson, and M. Eskenazi. An analysis of statistical models and features for reading difficulty prediction. In *Proc. of Workshop on innovative Use of NLP For Building Educational Applications*, pages 71–79, 2008.

[10] J. Kincaid, R. Fishburne, R. Rogers, and B. Chissom. Derivation of new readability formulas for navy enlisted personnel. In *Research Branch Report 8-75, Millington, TN: Naval Technical Training*, 1975.

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

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

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

[14] S. E. Petersen and M. Ostendorf. A machine learning approach to reading level assessment. In *CSL*, pages 89–106, 2009.

[15] E. Pitler and A. Nenkova. Revisiting readability: a unified framework for predicting text quality. In *EMNLP '08*.

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

[17] S. Schulz, K. Markó, E. Sbrissia, P. Nohama, and U. Hahn. Cognate mapping - a heuristic strategy for the semi-supervised acquisition of a spanish lexicon from a portuguese seed lexicon. In *COLING '04*.

[18] L. A. Sherman. *Analytics Of Literature: A Manual For The Objective Study Of English Prose And Poetry*. Ginn and Company, 1893.

[19] A. Stenner. Measuring reading comprehension with the lexile framework. In *American Conf. on Adolescent/Adult Literacy*, 1996.