Rodrigo Wilkens, Leonardo Zilio & Cédrick Fairon

Université catholique de Louvain, Louvain-la-Neuve, Belgium

{rodrigo.wilkens,leonardo.zilio,cedrick.fairon}@uclouvain.be

SW4ALL corpus: tagged and ready for searching

Bio data

Rodrigo Wilkens is a postdoctoral researcher at the Center for Natural Language Processing (CENTAL), Université catholique de Louvain (UCL), Belgium. He has a master's and PhD degree in Computer Science by the Federal University of Rio Grande do Sul (UFRGS), Brazil, and carried out a doctoral stay in the Massachusetts Institute of Technology (MIT), USA.

Leonardo Zilio is a postdoctoral researcher at the Center for Natural Language Processing (CENTAL), Université catholique de Louvain (UCL), Belgium, where he develops project SMILLE. He obtained his PhD in Linguistics at the Federal University of Rio Grande do Sul (UFRGS), Brazil, and carried out a one-year doctoral stay at the Laboratoire d'Informa-tique de Grenoble (LIG), France. He is also translator of German, English and Portuguese.

Cédrick Fairon is professor at the Université catholique de Louvain (UCL), Belgium, and director of the Center for Natural Language Processing (CENTAL). After a PhD in Computer Sciences at the University Paris 7, he was postdoctoral researcher at New York University (NYU, Linguistic Department) and worked as Associate Research Scientist at Educational Testing Service (Princeton, NJ) on automatic test generation.

Abstract

Learning a second language is a task that requires a good amount of time and dedication. Part of the process involves reading and writing texts in the target language, but the search for texts that are interesting for learners and pertain to their level is a time-consuming task. By focusing on the need for texts suited for different language learners, we present in this study a corpus classified by language proficiency level that allows the learner to observe ways of describing the same topic or content by using strategies from different proficiency levels. SW4ALL uses the alignments between the English Wikipedia and the Simple English Wikipedia and an annotation of language levels. We used an automatic approach for the organizing the corpus, followed by an analysis to sort out annotation errors. The resource contains 8,669 pairs of documents portraying different levels of language proficiency.

Conference paper

Introduction

Learning a second language is a process that requires exposition to texts, especially for the acquisition of vocabulary (Rott, 1999), and the task of retrieving texts that match learners' language level (or proficiency) can be achieved by using a corpus carefully designed for language learners, or by searching the web. Following these alternatives, systems can dig up texts aiming at finding those that are best suited to the language skills of a learner. Examples of these systems are REAP (Heilman et al., 2008), FLAIR (Chinkina, Kannan and Meurers, 2016), and READ-X (Miltsakaki and Troutt, 2007), which use the web as a corpus.

The use of the web allows for the interaction with a huge diversity of texts, but most of the web texts retrieved by search engines require a high language proficiency, even for native speakers (Vajjala and Meurers, 2013). On the other hand, the use of an off-line corpus ensures content quality, while hindering the search for different text topics.

This dichotomy of text sources enforces different types of restrictions on the systems. An alternative for trying to get the best of both approaches is to use the web as source of texts, but restricting it to trustful domains. This is similar to SourceFinder's method of downloading on-line newspapers and magazines and processing them as text sources (Passonneau et al., 2002; Sheehan, Kostin and Futagi, 2007). However, this approach makes it difficult to update the texts, because all content is stored off-line, and an update would require a rerun of the whole corpus compilation process. Another reliable text source that is adequate for language learners is a corpus of simplified texts, such as the Weekly Reader and the BBC Bitesize. This type of source generally represents an attempt to make texts more accessible for a non-native speaker. Wikipedia also has concerns on the comprehension abilities of its users, so that, for the English language, there is the Simple English Wikipedia, a simplified version, addressing the needs of natives with low literacy and also of learners of English. Resources such as these have a simplified version of a source article, serving as a facilitator for the communication of knowledge for those with less proficiency in the language, but they normally do not have information about to what extent the text is simplified. That is, they do not explain the simplification strategies applied or what is the target reader of each text. For language learning, this information is crucial to better inform the teacher or learner about which texts would be at an understandable level.

For instance, the Common European Framework of Reference for Languages (CEFR) classifies learners in six language levels (ranging from A1 to C2), while other frameworks use score ranges, like TOEFL and IELTS. Without this information about the language level, or some other information about the adequacy to a given target reader, language resources that present a simplified version are, per se, not well suited for language learners, because the applied simplification process may not help certain learners. As such, another classification with pedagogically relevant information is needed, i.e. a classification that takes into account the suitability of the text for learners, regarding the required proficiency.

This study aims at automatically determining the CEFR level of pairs of original and simplified texts, so that a corpus of aligned texts pertaining to different language levels can be used in a language learning framework. This would facilitate the comparison of text structures from different levels that describe the same content, while also allowing for the selection of topics of interest. So, we annotated language levels in an aligned version of the English Wikipedia[22] (EW) and Simple English Wikipedia[23] (SEW), and filtered pairs of texts that are associated to different levels, but that refer to the same topic or content. This process resulted in the *Simple Wikipedia for Aligned Language Learning* (SW4ALL)[24]. This resource could be employed by language teachers to select for their learners pairs of texts with an interesting topic that also present a contrast between two language levels in terms of writing.

Related Work

Aiming to reduce the time necessary to find texts that are interesting and at the learners' level, the SourceFinder (Passonneau et al., 2002; Sheehan, Kostin and Futagi, 2007) allows teachers to search for documents classified in different language levels by using keywords. One of the advantages of SourceFinder is the use of off-line texts, which allows the processing of the texts with several NLP tools without delay to the user.

22 https://wikipedia.org/.

23 https://simple.wikipedia.org/.

24 http://cental.uclouvain.be/resources/smalla_smille/sw4all/.

Using online documents, the REAP (Heilman et al., 2008), READ-X (Miltsakaki and Troutt, 2007), and LAWSE (Ott and Meurers, 2011) systems allow the users (learners or teachers) to search texts by using keywords and to filter them according to readability measures. Those systems identify the text readability by applying traditional readability scores (Flesch-Kincaid measure (Kincaid et al., 1975), and Gunning Fog Index (Gunning, 1952)) that are based on shallow cues (e.g. number of words per sentence and syllables per word). These measures have the advantage of a fast annotation process, but they are not accurate, and they require the users to deal with a score that may not be familiar to them.

Taking into account the pedagogical function of these systems, a major point is their ability to address documents that are readable by learners. However, text length-based readability scores consider only sentence length and word difficulty, ignoring factors such as cohesion (Bruce, Rubin and Starr, 1981). Recently, Xia, Kochmar and Briscoe (2016) compared syntactic and length-based features for text classification according to language level, and identified that adding syntactic features on top of length-based features improved the classification results, but using only length-based features presented a better result than the syntactic features alone.

All these systems focus on presenting a readable text, but some of them go beyond that, presenting exercises for supporting the learning activity in a more active way (e.g. REAP). In spite of all the effort to present readable and interesting texts to learners, those systems do not indicate how learners can improve their skills by using the indicated texts.

Regarding the representation of the texts in features, there is a huge variety of options in the literature. They can be grouped into 6 categories: length-based (e.g. word and sentence length), lexical (e.g. proportion of words in a list of easy words), morphological (e.g. part of speech), syntactic (e.g. presence of passive voice), semantic (e.g. word polysemy), and language model (e.g. n-gram model perplexity). Usually, a parser-based annotation of features follows the same process as the morphological annotation: simple counts of parser annotation. However, some studies, such as François and Fairon (2012) and Heilman et al. (2007), used information beyond parsed tags.

Methodology

For automatically determining pairs of texts that present good examples of different language levels, we trained a classifier and applied it to the aligned English Wikipedia (EW) and Simple English Wikipedia (SEW). In this section, we present first the alignment between the two Wikipedia versions, then we move on to the resources needed to build a classifier: a training corpus and a feature set.

Aligned Wikipedias

The Wikipedia is a collaborative encyclopedia with huge amounts of texts available in several languages. In English, there are two version, one that just focus on encyclopedic information, and the other that requires this information to be written in a simplified way. Comparing the vocabulary of the two encyclopedias, Coster and Kauchak (2011) identified that 96% of the words in the simple version are found in the other version, and 87% of the words in the normal version are found in the simple version. This overlap is also found at the n-gram level. Regarding the alignment of the Wikipedias, there are different versions, and in this paper we opted for the version organized by Kauchak (2013), in which the texts were, after a cleaning process, aligned both at the document and sentence level, resulting in 60K articles each. The difference in the number of sentences between the Wikipedia versions is partially because some articles from SEW contain only partial information. Indeed, the sentence level alignment presented by Kauchak (2013) was possible in only 28% of the sentences from the SEW (and in 4.25% of the sentences from EW).

Training Corpus

In this study, we used the EF-Cambridge Open Language Database (EFCAMDAT) (Geertzen, Alexopoulou and Korhonen 2013) as training corpus. The corpus is divided according to the

Common European Framework of Reference for languages (CEFR) (Verhelst et al. 2009), and each document has an evaluation score and an associated topic (e.g. introducing yourself by email). The EFCAMDAT corpus contains an unbalanced number of documents per level (e.g. 151 thousand documents for A2 and 23 thousand for B2), so we selected 9,000 random documents from each level, and also filtered out those documents that did not achieve an evaluation score higher than 90%, because, in these cases, learners' errors could have an impact on the machine learning approach (Pilán, Volodina and Zesch, 2016). The result of this process is a corpus of 40,946 documents (9,000 for levels A1, A2, B1, and B2, and 4,946 for C1[25]).

We decided to use a corpus of texts written by language learners, so that we are able to identify texts compatible with learners' productive skills. By applying to the Wikipedias a model that was trained on texts produced by learners, we expect that the structures in the retrieved texts will be familiar to the language learners, while also providing an authentic source of information, since the texts of the Wikipedias were written by native speakers.

Feature Annotation

The annotation process was three-fold: first the documents were automatically parsed with the Stanford Parser (Manning et al., 2014), and then a series of features were annotated, including grammatical structures that are pedagogically relevant, as conceived by Project SMILLE (Zilio and Fairon, 2017), and readability scores. The annotations were grouped in four categories, inspired by Xia, Kochmar and Briscoe (2016): length-based, morphological, syntactic, and readability.

In addition to these sets of features, we also considered the grouping of these morphological and syntactic features according to two criteria: CEFR level (e.g., A1, A2, etc.) in which they should be learned[26], which resulted in 5 grammatical and 5 word features (so that we had, for instance, A2 vocabulary and A2 grammar as feature); and type of grammatical structure (still respecting the CEFR levels division; for instance, connectives, which are learned on CEFR level B1, were put together in one set, but modals, which are learned in different levels, were separated in two sets). These sets of features were called pedagogical feature sets.

Classifier Performance

Corpus size is an important feature in machine learning. So, to evaluate the quality increase of our model in relation to corpus size, we performed ten experiments with varying corpus sizes, from 10% to 100% of the corpus. In each test, we used all feature sets and performed a ten-fold cross-validation with two algorithms (Simple Logistic and Random Forest)[27]. The model performance was evaluated in terms of precision, recall and f-measure. The f-measure increases an average of 0.15% for each increment of 10% in corpus size. However, in regard to statistical confidence, we identify a significant increase of f-measure only when the corpus is increased by at least 20% (1,590 instances), and no difference was observed in sizes larger than 40%. Despite the nonexistence of statistical difference in larger samples, they present a smaller standard deviation.

To the best of our knowledge, there are no studies addressing the classification of texts written by English learners. However, in the literature there are some studies that are similar to ours. For instance, Pilán, Volodina and Zesch (2016) address the same task, but using the MERLIN corpus (Wisniewski, 2013), which contains documents written in Czech, German, and Italian (80% of f-measure), and Xia, Kochmar and Briscoe (2016) employ a similar set of features, but using the Cambridge English Exams dataset, which is made up of text written by native speakers (80% of accuracy). Still, in an effort to establish a basis of

25 We used all documents that were scored over 90% C1 data, and we did not use the C2-level documents due to the low number of documents.

26 For allocating each structure to a given CEFR level, we used SMILLE's pedagogical model.

27 More details regarding the machine learning process can be found in Wilkens et al. (2018).

comparison, using corpus sizes that were similar to those studies, we achieved an F1 of 82% and an accuracy of 80%. If we consider the full corpus, we have an F1 of 84.7%.

Results

After the evaluation of our classifier, we applied it to the aligned version of the English Wikipedia (EW) and the Simple English Wikipedia (SEW), so that we could observe which pairs of texts are suitable for contrasting different CEFR levels of English. Based on the premises of the SEW that the texts should be simpler than the EW, they should at least be on the same CEFR level as their EW counterpart, so we considered that only pairs for which the system classified the SEW text as having a lower CEFR level than the EW were good for SW4ALL.

From more than 60 thousand pairs of texts, the system classified 9,222 pairs as having an SEW document that was classified as at least one level lower than its EW counterpart, which forms a more reliable set of pairs for language learning purposes. Another good chunk (10,225 pairs) was classified as having the same level in both Wikipedias. For these, we further looked at the distribution as a tie breaker. For instance, a pair in which both texts were classified as B1 was investigated to see if the distribution tended to show that the SEW text was easier than the text from the EW. This process identified 2,223 pairs of documents for which the SEW version tends to present a lower level. By adding these to the good sample, we got 11,445 pairs of texts in which the SEW document was deemed to present a lower level in relation to its EW counterpart.

To ensure the quality of the resource, we turned ourselves once again to the SEW assumptions, which indicate that texts should explain complex concepts to the user, while also splitting complex sentences, so as to form shorter, simpler sentences for the reader, but presenting the same content and with even longer texts (due to the explanations). With this information in mind, we cleaned from the corpus pairs in which the SEW text had a size (in number of words) of 90% or less than its EW counterpart, for the pair would almost certainly not present the same content, let alone a SEW version with more explanations[28]. This cleaning process removed 1,359 pairs of documents from the good sample, resulting in a subcorpus from the aligned EW-SEW containing 10,086 documents.

As a final step for ensuring the reliability of our classification, we clustered the distribution of probabilities for each level from the classifier using the k-means algorithm (Arthur and Vassilvitskii, 2007)[29], so as to distinguish how well our classified data matched the assumptions of the SEW. We divided the clusters into suited or non-suited, using purity as a measure of confidence, and so we were able to identify documents that possibly contain labeling errors. Considering only those documents that had a confidence score of more than 95%, SW4ALL consists of 6,394 pairs of documents (63% of all the documents that were considered suited), but, by relaxing this confidence to all of those above 85%, the size of the resource increases to 8,669 pairs of documents (86% of the documents that were considered suited), while maintaining a good confidence in the classification.

SW4ALL is available for search online and, by entering keywords, the user will have access to the different pairs of documents, with information regarding the level of each document and also the confidence that we have on the given pair, as explained above. The user can then click on the selected pair for accessing the actual web page content.

Final Remarks

This paper presents an interdisciplinary effort mixing Natural Language Processing and Second Language Acquisition to generate SW4ALL, a resource that focuses on the contrast of aligned texts that belong to different CEFR levels. This resource could be employed by

28 We did not restrict a maximum size, because it would be impractical to establish how much explicitations or sentence splitting would be too much.

29 The k value was set as 10.

teachers or learners as a means for comparing grammar, vocabulary and the general structure of texts in different levels.

A classification model was trained using an annotated version of the EFCAMDAT corpus, and the model was then applied for classifying pairs of aligned documents from the English Wikipedia (EW) and its simplified version, the Simple English Wikipedia (SEW). After further analysis, pairs of documents in which the document from the SEW were classified as having a lower level, or a tendency to have a lower level, were used in SW4ALL, resulting in a total of 8,669 pairs of documents. The resource is openly available for browsing and downloading, and displays information about the level of each document in the retrieved pairs, as well as the confidence that we have on them.

The pairs of documents present the same content or topic, so that SW4ALL can be a rich resource for aiding teachers and learners that wish to compare different linguistic strategies for writing a similar content, providing an interesting option for improving the learning of English as a second language.

Acknowledgements.
The authors would like to thank the Walloon Region (Projects BEWARE n. 1510637 and 1610378) for support, and Altissia International for research collaboration.

References

Arthur, D., & Vassilvitskii, S. (2007, January). k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms* (pp. 1027-1035). Society for Industrial and Applied Mathematics.

Bruce, B., Rubin, A., & Starr, K. (1981). Why readability formulas fail. *IEEE Transactions on Professional Communication*, (1), 50-52.

Chinkina, M., Kannan, M., & Meurers, D. (2016). Online information retrieval for language learning. *Proceedings of ACL-2016 System Demonstrations*, 7-12.

Coster, W., & Kauchak, D. (2011, June). Learning to simplify sentences using wikipedia. In Proceedings of the workshop on monolingual text-to-text generation (pp. 1-9). Association for Computational Linguistics.

François, T., & Fairon, C. (2012, July). An AI readability formula for French as a foreign language. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 466-477). Association for Computational Linguistics.

Geertzen, J., Alexopoulou, T., & Korhonen, A. (2013, October). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). In *Proceedings of the 31st Second Language Research Forum*. Somerville, MA: Cascadilla Proceedings Project.

Gunning, R. (1952). *The technique of clear writing*.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Human Language Technologies 2007*: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference (pp. 460-467).

Heilman, M., Zhao, L., Pino, J., & Eskenazi, M. (2008, June). Retrieval of reading materials for vocabulary and reading practice. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 80-88). Association for Computational Linguistics.

Kauchak, D. (2013). Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st annual meeting of the association for computational linguistics* (volume 1: Long papers) (Vol. 1, pp. 1537-1546).

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel (No. RBR-8-75). *Naval Technical Training Command Millington TN Research Branch*.

Miltsakaki, E., & Troutt, A. (2007, October). Read-x: Automatic evaluation of reading difficulty of web text. In E-Learn: World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education (pp. 7280-7286). *Association for the Advancement of Computing in Education* (AACE).

Ott, N., & Meurers, D. (2011). Information retrieval for education: Making search engines language aware. *Themes in Science and Technology Education*, 3(1-2), 9-30.

Passonneau, R., Hemat, L., Plante, J., & Sheehan, K. M. (2002). Electronic Sources as Input to GRE® Reading Comprehension Item Development: SourceFinder Prototype Evaluation. *ETS Research Report Series*, 2002(1).

Pilán, I., Volodina, E., & Zesch, T. (2016). Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*: Technical Papers (pp. 2101-2111).

Rott, S. (1999). The effect of exposure frequency on intermediate language learners'incidental vocabulary acquisition and retention through reading. *Studies in second language acquisition*, 21(4), 589-619.

Sheehan, K. M., Kostin, I., & Futagi, Y. (2007). SourceFinder: A construct-driven approach for locating appropriately targeted reading comprehension source texts. In *Workshop on Speech and Language Technology in Education*.

Vajjala, S., & Meurers, D. (2013). On the applicability of readability models to web texts. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations* (pp. 59-68).

Verhelst, N., Van Avermaet, P., Takala, S., Figueras, N., & North, B. (2009). *Common European Framework of Reference for Languages: learning, teaching, assessment*. Cambridge University Press.

Wisniewski, K., Schöne, K., Nicolas, L., Vettori, C., Boyd, A., Meurers, D., ... & Hana, J. (2013). MERLIN: An online trilingual learner corpus empirically grounding the European reference levels in authentic learner data. In *ICT for Language Learning 2013*, Conference Proceedings, Florence, Italy. Libreriauniversitaria. it Edizioni.

Xia, M., Kochmar, E., & Briscoe, T. (2016). Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 12-22).

Zilio, L., & Fairon, C. (2017, July). Adaptive system for language learning. In *Advanced Learning Technologies (ICALT)*, 2017 IEEE 17th International Conference on Advanced Learning Technologies (pp. 47-49).

Wilkens, R., Zilio, L., & Fairon, C. (2018). SW4ALL: a CEFR-Classified and Aligned Corpus for Language Learning. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (pp. 365-370).