A survey for readability assessment

Ion Madrazo Azpiazu

Introduction

This paper presents a survey on automatic readability assessment techniques. The survey, starts by making a summary of the most relevant papers in this area and ends by describing some interesting AI methods that could be used in the area.

The final aim of this survey is to support a project of development of a readability assessment tool for basque. With that goal in mind, a survey of the most common techniques in the area has been carried, making an special emphasis in not fully studied languages regarding Natural language processing, such as basque, arabic, or italian. Finally, some papers related to AI algorithms are described at the very end. Those are papers that contain algorithms that will be used in the final project.

Simple or Complex? Assessing the readability of Basque Texts (2014)

Itziar Gonzalez-Dios, Maria Jesus Aranzabe, Arantza Diaz de Ilarraza, Haritz Salaberri

This paper presents the only existing readability assessment system for basque language. Being Basque a minority language, and a specially different one due to its pre-indoeuropean origin, very little research has been done compared to other popular languages.

This paper tries to get rid of the peculiarities of basque and presents a very simplified version of a readability assessment system. The system is able to predict two class values given a natural language text, “simple” and “complex”. Simple, refers to a text that is considered apt for a 12 year child. Complex refers to a highly technical text oriented for a graduate student or higher.

The features used, are ratios between common NLP features and text lengths. Features such as, verb, adjective or noun count per text. Those features are used to ran a model based on support vector machines and predict the results of incoming texts.

The main critique for this paper, is that the system is only able to predict only two labels, while most of the systems for other languages are commonly able to predict 4 or 6. No pure syntactic features are used either, features that suppose an improvement in other systems in the area. However, given the limitations of resources for Basque, the paper supposes a big step forward in the area of Basque readability assessment.

Chinese Readability assessment using Tf-Idf and SVM (2011)

Yaw-Huei Chen, Yi-Han Tsai, Yu-Ta Chen

This paper presents a readability assessment tool for chinese language. Rather than using common linguistic features, the presented system makes uses of a unigram model. The frequencies of terms are used as features for documents. Those frequencies are used for training a SVM model and predict the readability level of new documents.

The system is able to predict the readability level for 3 different grades of elementary school. For each grade a different binary classifier is trained, using the documents of the grade as positive examples and the other documents as negative.

The most important step of the system is the feature/term selection step. The system needs to figure out which terms are important for each grade, since the whole set of terms is too big to be used as features. For this, a mutual information measure is used. This metric, is able to reveal the importance of a term to a certain class. This is done by computing how frequently a term appear in a document of the given class, and how frequently it does not appear in other documents. In some means, it works similarly to a Pearson correlation measure.

For each of the selected term its Tf-Idf value is computed to form a vector of frequencies. This vector is the input vector for the classification model. The Tf-idf value is nothing more than a frequency value that is normalized against the usual frequency of the term. This way word that tend to appear very often such as “and” or “the, get their value lowered and are therefore, not selected as features.

My main critique for this paper is that the system is not topic independent. Since the models are trained in certain books for elementary school, the have the term vocabulary of those certain subjects. So , if the system is tested against books from other subjects, the system should not perform as good. This does not happen on other systems that use common NLP features, since those features show how the text is structured and are not reliant on the vocabulary of the text itself.

Towards the Development of an Automatic Readability Measurements for Arabic Language (2008)

Amani A. Al-Ajlan , Hend S. Al-Khalifa, AbdulMalik S. Al-Salman

This paper, rather than presenting a full readability assessment tool, shows the importance of two very simple features for the task. It shows that the performance of a SVM model could be good just by learning over those two features.

The two features are nothing more that two different ratios. The first one is the character per word ratio. The hypothesis behind this feature is that the more technical and sophisticated a text is, the longer the words are. Surprisingly, the feature has a very strong correlation with the level of readability of the texts.

The second feature is the words per sentence ratio. The hypothesis behind this is similar, the more complex a text is, the longer sentences it has. This ratio is highly correlated with the class value too.

My main critique to the paper is its lack of results. They show how well correlated are those features, and even how their system would work over two certain instances, but, the do not show it in a significantly big dataset.

A Language Modeling Approach to Predicting Reading Difficulty

Kevyn Collins-Thompson, Jamie Callan

This paper follows a similar approach to “Chinese Readability assessment using Tf-Idf and SVM” in the sense that it uses a unigram based language model for predicting the readability level of texts. However, it solves some issues presented in that paper, issues such as topic dependance and works better that other similar systems in short passages containing less that 10 words.

The system is able to predict the readability for 12 different grades of reading. The language model is defined as by the terms that appear more frequently for each of the readability classes. As they state, the problem of the topic dependance can be mostly solved by deleting the top most frequent terms, since those are the most topic specific terms. I find this argument quite hand wavy, since the do not give any proof of it, it is just a supposition made by visual means.

The model the system uses for learning is a modified naive bayes model. The model has been modified to not treat the features as fully independent. As the authors state, there probably is a high dependence between terms that appear frequently in adjacent readability

levels. They insert this dependence to the naive bayes model in order to increase the performance of the system. This dependence is inserted using an interpolation function for smoothing the term frequencies all over the different class values.

Automatic Readability Assessment for People with Intellectual Disabilities

Lijun Feng

This paper aims towards the development of a system that is capable of simplifying texts, in order to be more readable for people with intellectual disabilities. The simplification process has not been developed yet, but the readability assessment system has. This system will be a submodule of the final simplification system.

The system follow a common SVM approach for the learning model. However, its interest lays in the features that have been developed for the model. The system makes uses of a high variety of NLP features. They range from traditional lexical to discourse level features, passing form some syntactic features. The discourse features, are features that show how the text is structured, they are based mostly in connectors between sentences and paragraphs.

The results show that the syntactic and discourse features they introduced outperform the baseline shallow traditional features.

As future work, they state that the results that the system have achieve will be tested in a real environment with people with intellectual disabilities, to show the real usefulness of the system. Even if these results would be interesting to analyzed, and are what attracted me to read the paper, to the best of my knowledge, these results have not been published yet.

READ-IT: Assessing Readability of Italian Texts with a View to Text Simplification (2011)

Felice Dell'Orletta, Simonetta Montemagni, Giulia Venturi

This paper presents a readability assessment system for Italian. However, since the aim of their project is focused on text simplification, their system is strongly biased for that tasks.

The text simplification system the authors are working is based on sentences. Therefore, the aim of this readability system is to predict whether a sentence needs simplification or not.

Given that fact, the system will only need to predict two class values, whether the sentence is difficult or simple.

As the system work in sentence level, all the features appearing in other papers that go further that sentence level are discarded. So the feature that this system uses for learning are mainly lexical and syntactical ones. They use some interesting features such as, syntactic dependencies, which are novel to the area due to their high computing cost.
Even if the system is focused to predicting sentence readability, the authors have made an experiment to show how their system would work in text level. For this, every sentence of the text is first predicted, to finally, calculate the overall prediction of the full text, given the percentage of difficult and easy sentences.

As a critique, I would like to see a comparison between their systems starting from sentence level and their same system by starting directly at text level. This could support or not their statement that the readability systems would work better by calculating readabilities at sentences level. I think the idea is quite good, and I am mostly convinced that it would work, but I need some evidence for fully believing it.

A comparison of Features for Automatic Readability Assessment (2010)

Lijun Feng, Martin Jansche, Matt Huenerfauth, Noemie Elhadad

This paper makes a comparison of the currently most used features for automatic readability assessment. The comparisons are made using each of the features with a SVM classifier. Most of the features are features already shown in this survey, therefore i will center this description on two features that have not been mentioned yet .

On the one hand, the paper presents a feature called coreference inference based feature. This feature makes uses of the commonly used coreference resolution algorithms of NLP. Those, basically are able to make links between pronouns of text, to show which of them are directly related. These links are used to generate a graph which is then analyzed to determine its complexity and used as feature.

On the other hand, they present another novel feature based in named entities. A named entity is a set of words that refers to a certain entity, such as Apple, Microsoft or Obama. A graph is generated by determining the relations between entities in the text. This graphs is then used to determine how coherent is the text by seen how big the jumps between entities are. E.g. if a sentence talks about, microsoft and the next sentence talks about apple, the jump is considered coherent, because those two entities are somehow related. However if the jump is from microsoft to coca cola, the jump is considered less coherent.

Personally I find, this paper very advisable for someone that is starting in the area of readability assessment. It gives a good general idea of what have been done in the area.

Readability Assessment for Text Simplification (2010)

Sandra Aluisio, Lucia Specia, Caroline Gasperin, Carolina Scarton

The aim of the system presented in this paper is different from the ai of other readability assessment systems. Rather than creating a system for the very end user, this system has been created in order to evaluate a text simplification tool.

The aim of the system is to predict whether the input text is an original complex text, a text that has been slightly simplified for average people's understanding, or a text that has been strongly simplified for people with understanding difficulties.

The system makes use of a high variety of NLP features extracted from texts. Syntactic features seem to have the highest impact in the performance of the system. In addition, they present a novel feature to the area, which describes the ambiguity of the text. This feature counts the number of senses each of the term in the text can have in order to detect how complex is for the reader to understand the text. The higher the number of senses in text, the higher the different interpretations a text can have, and therefore the more difficult the text is to understand.

The paper reports a 0.904 F-score, which is incredibly high in the area. This leads me to think that the task they are resolving is easier that the common readability assessment task, since the input texts are generated by a machine. I would be interesting to see how their algorithm performs in a common readability assessment environment, with texts written by humans.

An "AI readability" formula for French as foreign language (2012)

Thomas Francois, Cedrick Fairon

This paper presents a readability assessment tool for French as foreign language. The systems is based on a supervised classifier model. The features they present are highly diverse, being the following one, the most uncommon ones:

They present a novel feature called "conceptual density", which considers the average number of different arguments per sentence. The more arguments a sentence have, the more

complex its reasoning is and therefore the more complex the text is in overall. These propositions are extracted using 35 different rules, they have developed.

Another novel feature they present is the “lexical diversity” feature. This feature considers that the more high level a text is the richer vocabulary it has. Therefore, this feature takes into account the ratios of repeated words in the text.

Lastly, they present an approach that uses a list of theoretically easy words. These words are taken from children textboxes. The more word of a text appear in that list, the simpler the text is considered.

The system has been tested with different learning models, being the SVM model the more effective one. The results are given in a novel way to the area. Instead of using the usual accuracy measure, they make use of a measure called adjacent accuracy, which counts as correct the errors with distance 1.

A Simple Approach to ordinal Classification (2001)

Eibe Frank and Mark Hall

Machine learning problems commonly assume that there is not order in the class values. However, many real life problems do exhibit a natural order in their class values. Therefore, this paper focuses its work in developing a method to take benefit of that natural order.

Historically, these problems have been treated as regression problems, even if the class values where discrete. Values were treated as continuous values and learned in a regression model, to finally be discretized again. This paper takes a different approach to it, it show a technique that can be applied in a classification context.

In summary, the algorithm they present takes the discrete values and make groups of then using inequalities. E.g. The problem of determining a temperature value using the discrete values “Hot” , “mild” and “Cool” is converted to a new problem that uses inequality groups, “bigger than cool” and “bigger than mild”. Several binary classifiers are trained to learn and predict each of those new classes. Each binary classifier outputs a probability value. Those values are used to determine the most probable value for the class.

The evaluation of the system is in general solid to demonstrate that their system outperforms usual classifiers that do not take order into account. However, some comparison between the common regression approach and their approach would be interesting to see.

Ordinal Regression by Extended Binary Classification (2006)

Ling Li and Hsuan-Tien Lin

This paper, tries to solve the same problem as “A Simple Approach to ordinal Classification”. So the objective is to develop a learner that takes advantage of the ordinality in the class values. However the approach taken is different from the mentioned paper.

Firstly the label values are transformed from single values to inequalities, in the same way as in the mentioned paper. So if the class contains 5 values (1,2,3,4,5) the converted values are 4 (greater than 1, greater than 2, greater than 3, greater than 4).

Secondly, a single binary classifier is trained. This classifier has an extra value as a input apart from the features that the problem already contained. This new feature shows the inequality that the current example is testing (greater than 1, greater than 2, greater than 3, greater than 4). The output of the classifier will show whether this inequality is correct or no (true,false) for the given features.

In the classification phase, the instance is classified for all the inequalities. For each one, a true or false value is determined. The number of true outputs are summed in order to obtain the final result. E.g.: If the classifier has outputted 3 trues (greater than 1 = true, greater than 2 = true, greater than 3, greater than 4 = true) the output will be 3. Furthermore, the percentage of true below the highest true value can be used as a confidence value. The previous instance would have a 75% confidence value.

As a critique to the paper, I missed some comparison between their method and other papers stated in their related work. Even if they claim that their method is more generalizable that the others, no performance comparison are shown.

Making better use of global discretization

Eibe Frank and Ian H. Witten

Before applying learning algorithm, it is a common approach to discretize the datasets. The reasons for this can be that the learning algorithm cannot handle continuous values or just that the results are better using discretization. However this discretization process usually lead to a loss of information regarding ordinality in data. As ordinal features are discretized, any information in its order is lost.

This paper, present a new approach for discretization that conserves some order information in ordinal continuous features. The algorithm works as the following:

For a feature with n values, n-1 binary features are created. Each of those new features consist of a inequality that splits the values of the feature into two. E.g: If the values of the feature are {v1,v2,v3,v4), the following three discrete and binary features are created {value > v1, value >v2, value>v2}.

The algorithm presented significantly increases the performance of algorithms such as C4.5, compared to the common discretization methods.

Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning

Usama M Fayyad, Keki B.Irani

As previously stated, discretization is an important step of classification, just because the learning algorithms cannot handle continuous values or just because discretizing data leads the system to better results. This paper centers its work in trying to develop a discretization method that will make the work of the classification models easier.
The discretization method that is presented, is based on the metric of entropy from information theory. The method they propose splits the attribute values in multiple interval in order to minimized the sum of entropies regarding the class values. The algorithm they present works in a recursive way, and uses an heuristic to accept or reject a partition made in every step.

The heuristic is composed of two main features. On the one hand the heuristic prefers cutting points that make the overall entropy small. However on the other hand the heuristic downvotes having too much cutting points. This second heuristic is used because otherwise, the algorithm would always choose a discretization were each value is alone in its own bin.

The method looks to make significant improvement regarding the accuracy of classification models.

Discretization based on Entropy and Multiple Scanning

Jerzy W. Grzymala-Busse

This paper presents an iterative discretization approach called dominant attribute algorithm. The algorithm is based on entropy and information gain values.

For each step of the process, only one feature and one cutting point is selected. That cutting point is the one that maximized the information gain ratio of all the attributes regarding the class. Every time a cutting point is selected, the entropy is reduced and one new loop starts.

The loop ends, in two conditions: If the entropy can no longer be reduced. If every value of the class can already be distinguished by the already discretized attribute values.

The results show that the system outperforms commonly used, equal interval width and equal frequency per interval methods.

Practical Feature Subset Selection for Machine Learning

Mark A. Hall, LLoyd A. Smith

This paper presents a novel approach for feature subset selection. Feature selection, is one important step in the development of a classification. The good selection of the features of the problem can lead to better performance of the classification system in means of both accuracy and computation or storage efficiency.

This paper presents an algorithm that aims to select a subset of features for the a problem, trying to conserve/improve accuracy, lowering the number of features for the problem. To select an optimal subset from a set is a NP hard problem, so the algorithms cannot assure to find the optimal subset. Therefore, the algorithm uses traditional space search algorithms such as best first or hill climbing.

These algorithms make use of a heuristic to determine the goodness of a subset, and move through the space. The heuristic used by this system is called CFS (Correlation based feature selection) and is based in a correlation measure. A subset is better, the more correlated its features are with the class and the less correlated are the features between them. This heuristic tries to get the features which have the most valuable information, while avoiding redundancies in data.

The algorithm increases the performance of the learning systems in most cases and is fast enough to be used with a relatively big amount of features. The authors state that it can their algorithm took 8 minutes to execute in a 36 feature dataset, while another state of the art feature selector took around 8 days. So the increase in performance is quite significant.

Conclusions and future work

This survey has shown that, the area of the readability assessment is not a very diverse area regarding the learning models used. Most systems make use of common classification algorithms such as SVM. As a possible work regarding this, it would be interesting to see how these learning models could be enhanced, e.g. using an ordinal classification method.

Most of the work in this area is related to feature engineering, showing a lack of agreement between authors regarding which features work better.This may be due to the fact that the systems presented are oriented to different languages, and every language has its way to make texts easier or more difficult. An interesting study in the area would be one that made a comparison of a exactly same system in different languages, in order to see whether the features affected in the same way to the performance, unregarding of the language.

After seeing, the various features used in the area, I detected a lack of use of various higher level features, such as syntactic dependencies, or features regarding phenomena such as synonymy or followability of texts.

In conclusion, the of readability assessment seem to be an area yet to be deeply explored and that will surely suppose challenging problems for both artificial intelligence and natural language processing.